ITEM SAMPLING IN EDUCATIONAL RESEARCH

T. R. Husek and Ken Sirotnik

CSEIP Occasional Report No. 2, December 1967
University of California, Los Angeles

The research and development reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education under the provisions of the Cooperative Research Program.

ABSTRACT

This paper introduces the concept of item sampling by considering the possible cases that can arise when making inferences from a sample of examinees and/or test items to a population of examinees and test items. Item sampling is the case in which both examinees and items have been randomly sampled from a population of examinees and items. Formulas have been worked out such that the data from the item sample can be used to estimate the mean and variance of the population. Computational examples are provided in the appendix to illustrate the cases discussed.

Several further cases involving item samples are introduced: (1) single item samples in a one-way anova with two levels, (2) multiple item sampling for increased accuracy in inference, and (3) multiple item sampling in a one-way anova with two levels plus extentions to a two-factor design.

Recommendations based on examinee score and item difficulty variance are made for deciding on the number of subjects vs. the number of items to sample from the population. Certain advantages as well as cautions pertaining to item sampling are discussed, and possible used of item sampling in educational research are proposed.

ITEM SAMPLING IN EDUCATIONAL RESEARCH

Most educational researchers collect empirical data as part of their endeavors. Usually, the data are the responses of a number of subjects to a collection of items. It is hoped that the data will shed light on the substantive research questions. Sometimes, the responses of the individual students to the individual items are the data of interest, while at other times, the responses of the individual students to the total pool of items are desired. Occasionally, some indices of the responses of the total group of subjects to the entire test are examined, for example the mean and standard deviation of the test scores.

In none of the above situations does the researcher go beyond the data which has been collected; his aim is merely to describe in some manner the responses of the subjects who were tested to the items that were used. The analysis of the data ranges from a simple report of the responses of the students to the computation of summary statistics on the performance of the group. Statisticians refer to the collection of possible analyses as descriptive statistics--that is, descriptive statistics are used when the researcher does not want to make any generalizations beyond the data which have actually been collected. Often, however, the educational researcher wants to make statements about larger groups of data, which have not been collected but which might have been collected. For example, a population of subjects is postulated, a sample of subjects is drawn from the population, the sample of

subjects is given a test, and the results of the measurement procedure are analyzed not so much with a view to describing the sample results but with the aim of making statements about the inferred performance of the total population of subjects on the test. The performance of the sample of subjects is, in itself, of secondary importance; the primary interest is in making inferences from the sample results to the population. The targets of investigation are not the mean and standard deviation of the sample, but the estimated mean and standard deviation of the population.

For every descriptive statistic that might be computed for a sample of people, there is a corresponding parameter for the population of people from which the sample was drawn, and there is a large collection of statistical procedures for making inferences about a population of people based on the data collected from a sample of people. There is, moreover, a growing tendency to use the term "statistical inference" to describe the activity of using data on a sample of people to make inferences about a population of people.

It is also possible to make inferences about a population of items---that is, if a group of subjects is given a collection of items, it is possible to treat the items as a sample from a population of items and the tested subjects as the only subjects of interest. In this case, the inference is from the performance of the subjects on the sample of items to the performance of the same subjects on a population of items. "Psychometric inference" is the term which is used to describe this activity. It is

important to keep statistical inference, (where the inference is from a sample of people taking a fixed set of items to a population of people and the same items), separate from psychometric inference where the inference is from a fixed number of subjects and a sample of items to the same subjects and a population of items.

Recently, attempts have been made to develop systematic procedures for the case in which both people and items are sampled and where the inference is from a sample of people taking a sample of items to the performance of a population of subjects taking a population of items. To put the point more clearly, a random sample of subjects respond to a random sample of items, and statistical inferences are made simultaneously with psychometric inferences. Most of the work has been done by Frederic Lord. In his publications, Lord uses the term "item sampling" to cover situations when items are sampled. Since items are sampled in the case we have labeled "psychometric inference" and also in the case where both subjects and items are sampled, we have coined a new term "statistical-psychometric inference" to use in the latter situation. This is a paper about "item sampling," but we have chosen to divide the general topic into two more specific cases. It should be understood from the outset that our treatment of "statistical-psychometric inference" is not different from Lord's presentations on "item sampling," and is often based on Lord's analyses.

The three types of inference described above should be kept separate from another kind of inference, the non-statistical inference from a population that was randomly sampled to a larger

target population that was not randomly sampled. In "statistical inference," in "psychometric inference," and in "statistical-psychometric inference," the inference is from a random sample of items and/or people to a population of people and/or items. The major strength of the inferences is that probability theory can be used to describe their accuracy and use of probability requires that random sampling be used. In much educational research, the population that is randomly sampled is not really the major target population, and a non-statistical inference is made only after the statistical analysis is finished. The following situation is an example: A researcher has access to a school district. He obtains a random sample of students from the school district, and obtains data on the random sample of subjects. He uses statistical inference procedures to generalize from the performance of the sample to the population of subjects in the school district. But he really wants to generalize to all the schools in the area, perhaps in the county, perhaps in the state. These generalizations are made all the time and they are not necessarily restricted to educational research. Yet, they are not statistical in nature---they are not really based on probability but, rather, on the judgment of the researcher that he will not do violence to his conclusions if he makes a larger generalization than is completely warranted by his data gathering procedures.

These, then, are the major uses to which the educational researcher uses data he collects. He might just desire to describe the available data; he might want to make a "statistical inference"; he might want to make a "psychometric inference"; he might want

to make a "psychometric-statistical inference"; and he might want to make an "extra-statistical inference." The various inferences are pictured in Figure 1.

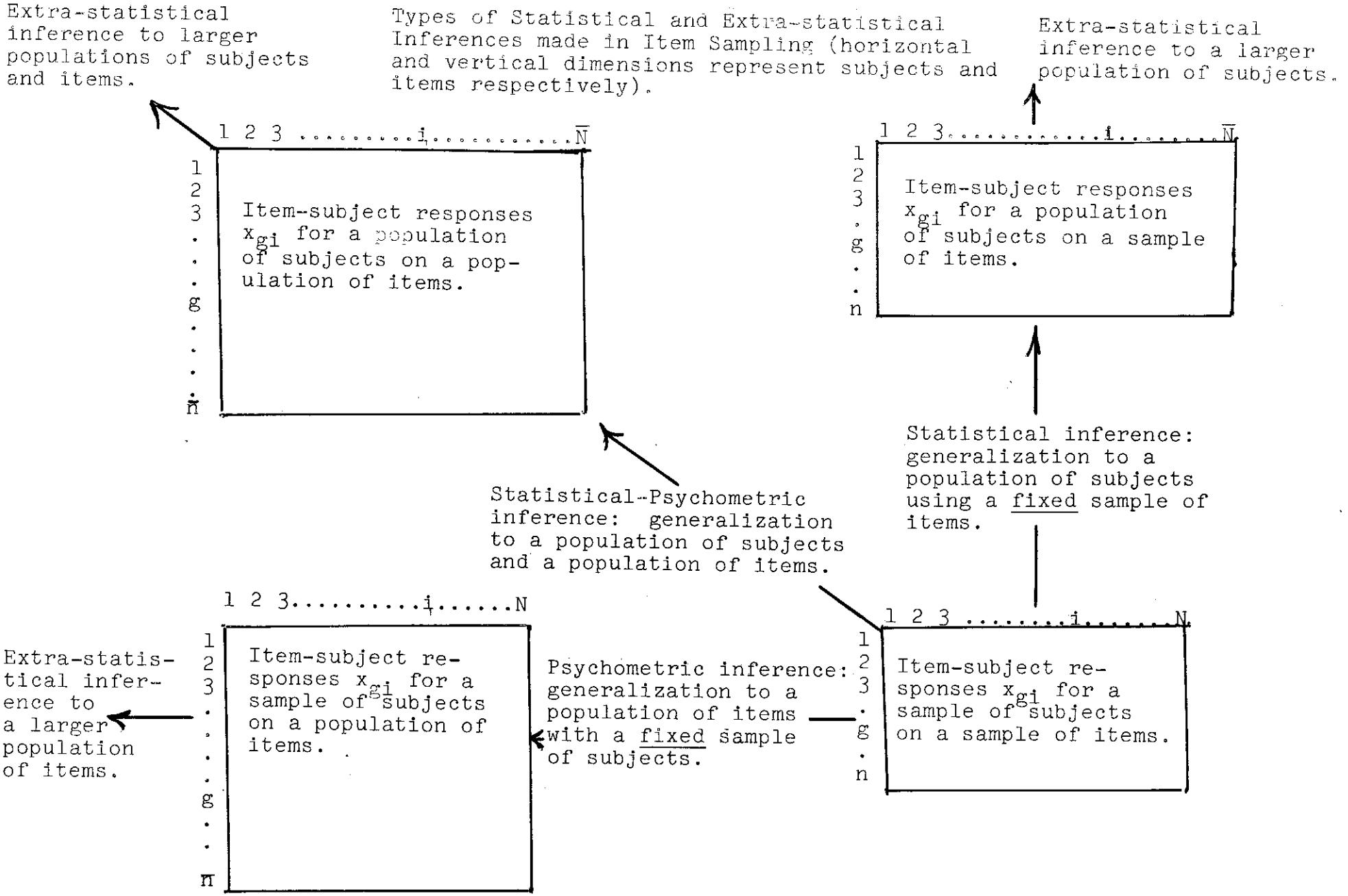
Hopefully, the foregoing has achieved its purpose: it has served as an introduction, placing item sampling in perspective.

If the reader turns to the literature, he will find that little has been written about item sampling. The other kinds of inference have been widely used and are extensively described in statistics books and philosophy of science texts. "Statistical-psychometric inference," however, is a new field, and Frederic Lord has been primarily responsible for its development. Lord's interest has been that of a psychometrician, and he treats item sampling both as a basis for test theory and as a novel technique for gathering data in some situations. We are not treating item sampling as a basis for test theory in this paper; our interest, instead, is in trying to explain item sampling as a new approach to gathering educational data---an approach which appears in many cases to be far superior to existing techniques.

In the next section of this paper, six general cases for educational research will be presented. We hope that these six cases will include most possible uses of item sampling. Not every case will, strictly speaking, include a different use of item sampling, but some seemingly extraneous material is included for clarity of presentation. At the conclusion of the paper, some general statements about item sampling will be made. For the time being however, let us make only one assertion: while item

FIGURE 1

Types of Statistical and Extra-statistical Inferences made in Item Sampling (horizontal and vertical dimensions represent subjects and items respectively).



for the various situations (viz., both item and subject populations infinite, both finite, or one finite and the other infinite) are rather complicated; for all practical purposes, the usual formulas for statistical inference will serve adequately. Further discussion on this latter point will follow.

The following cases begin with the most simple situations and become more complex. The presentations of the simpler uses of item sampling are hopefully clear; the more complex cases are built on the earlier material. The reader is requested to be patient if the descriptions at first do not seem appropriate for his needs for the later cases make much more efficient use of item sampling. The reader may also feel that statements about optimal sample sizes are not sufficiently specific. They are not specific because they are not known. However, as discussed later in the section of general advantages, the exact specification of the optimal procedure is not always necessary, especially in the situations where the alternative to an adequate, but perhaps non-optimal, procedure is no research at all.

CASE I: Statistical or Psychometric Inference: Simple Item or People Sampling

If a sample of people is drawn from a population of subjects and is given a fixed set of items, simple formulas available in most elementary statistics texts are adequate to provide estimates of the mean and variance for the population. The mean of the sample is an unbiased estimator of the mean of the population and the sample variance (multiplied by a correction factor) is an

unbiased estimate of the population variance. Written in relative score notation (see Appendix 2 for all notation),

$$(1) \qquad \bar{y} = \frac{\sum_{i}^{N} y_i}{N} = \mu_y$$

where

$$(2) \qquad \mu_y = \frac{\sum_{i}^{\bar{N}} y_i}{\bar{N}}$$

$$(3) \qquad \frac{N}{N-1} \; s_y^2 = \sum_{i}^{N} (y_i - \bar{y})^2 = \hat{\sigma}_y^2$$

where

$$(4) \qquad \sigma_y^2 = \sum_{i}^{\bar{N}} (y_i - \bar{y})^2 \; / \bar{N}.$$

If items (and not subjects) are sampled, the procedures are the same. The formulas, however, look a little different.

$$(5) \qquad \bar{p} = \frac{\sum_{g}^{n} p_g}{n} = \hat{\mu}_p$$

where

$$(6) \qquad \mu_p = \frac{\sum_{g}^{\bar{n}} p_g}{\bar{n}}$$

(7)

$$\frac{n}{n-1} s_p^2 = \sum_g^n \frac{(p_g - \bar{p})^2}{n-1} = \hat{\sigma}_p^2$$

where

(8)

$$\sigma_p^2 = \frac{\sum_g^{\bar{n}} (p_g - \bar{p})^2}{\bar{n}}.$$

It can be shown quite easily that

(9)

$$\bar{y} = \bar{p} = \hat{\mu}_y = \hat{\mu}_p$$

(see appendix 1, case 1, note). Case I is the simplest case, and, as previously indicated, the inferential procedures for Case I are given in most elementary statistics texts. (Simple examples using Case I are outlined in Appendix 1.)

CASE II: Simple Statistical-Psychometric Inference: Simple Item and People Sampling

In this case, a sample of people is given a sample of items, and the obtained data is used to generate estimates of the performance of the population of people on the population of items. The formulas for estimating the mean and variance of the population of subjects taking the population of items are either of formulas (1) or (5) for the mean and

(10)

$$\hat{\sigma}_y^2 = \frac{N(\bar{N}-1)}{\bar{n}(n-1)\bar{N}(N-1)} \left[n(\bar{n}-1) s_y^2 - (\bar{n}-n) \left\{ \bar{y}(1-\bar{y}) - s_p^2 \right\} \right]$$

for the variance. Notice the difference between the estimates of σ_y^2 as given by (3) and by (10). Formula (10) is far more complex

for 2 reasons: It assumes (a) that the item and subject populations are both finite and (b) that the inference must now not only be made to a population of examinees, but to a population of items. The reader should recall that in the section preceeding Case I it was recommended that both item and subject samples be treated as being infinitely large. This would modify (10) as follows:

(11)

$$\hat{\sigma}_y^2 = \frac{N}{(n-1)(N-1)}\left[ns_y^2 - \{\bar{y}(1-\bar{y}) - s_p^2\}\right].$$

It was also stated that the usual formulas for statistical inference can be used in all item sampling situations. In other words, although formula (11) gives the exact estimate of σ_y^2 (assuming infinite populations), formula (3) will also give this estimate, not quite as accurately, but adequately and simply. Computationally, then, we can treat Case II in the same manner as we treated Case I; the statistical inferences we make, however, are to two populations instead of only one. (The difference between the various formulas is illustrated with a computational example in Appendix 1, Case II, note.)

At this point, item sampling procedures provide the investigator with a number of alternatives which are not ordinarily available to him; for, if the researcher has a given population of subjects and a given population of items, he must make some decision about how many items and how many subjects to sample. He might give all the subjects all of the items; he might give all the subjects one of the items; he might give one subject all the items; or, he might give some of the items to some of the subjects. The question

is which of the possible data gathering procedures is most efficient. To help explicate the issue, an example will be presented. It will also be used to illustrate later points.

Let us consider a large high school which has been given some funds to establish a remedial reading program for some of its students. Six hundred students are eligible for the program, and 300 of them are selected at random and placed into the program. The school staff develops a set of 100 items based on the instructional objectives that they develop for the special program. The school administrators would like to make some statements about the effectiveness of the program. It is decided that because of time and cost considerations, only 900 observations are feasible. (An observation is defined as the response of one student to one item.) The number 900 is not magical. There are usually some restraints on the amount of data that can be collected, often based on administrative constraints of cost problems. The number 900 was chosen just to make the example more specific.

Thus, for Case II, there is a given population of 300 subjects, a given population of 100 items and constraints that only 900 observations are possible. What's more, in Case II, the researcher is restricted to giving one sample of items to one sample of subjects. However, there are still a large number of alternatives: three items could be given to 300 subjects; 100 items could be given to a sample of nine subjects; and of course, there are many possibilities between these extremes.

The following are the criteria for selecting data gathering procedures. First, it is necessary to obtain some information about the relative homogeneity of the population of subjects and of the population of items. More specifically, it is important to have some idea of the variance of relative observed scores of the population of subjects (σ_y^2) and the variance of the item difficulties in the population of items (σ_p^2). If the items are homogeneous in this special sense and if the subjects are quite variable, it is advantageous to take a small sample of items and give these items to a large sample of subjects. If the items are variable and the subjects are homogeneous, it is advantageous to give a large sample of items to fewer subjects. It is difficult to specify the exact number of items and subjects to sample (the reader is referred to Lord (1965) for a more complete treatment of the topic), but the criteria are very useful as guidelines. It is not necessary to actually know the two variances mentioned above, for available guesses would probably be quite useful.

In the example of the high school remedial program, it should not be difficult to estimate the variability of the subjects and the variability of the item difficulties. If the items are well tied to the objectives of the program, it is probable that the items will be less variable than the subjects and a few items should be given to a larger group of subjects. The reader who is concerned about the restraints imposed on the example is requested to be patient because Case II, in itself, is not an efficient use of item sampling. (A simple example using Case II is outlined in Appendix 1.)

CASE III: Statistical-Psychometric Inference With Two Subject Samples and One Item Sample; Simple Hypothesis Testing

In this case, the researcher wishes to perform an experiment and test the hypothesis that the mean score of one population of people is the same as the mean score of another population---that is, the researcher has two populations of subjects and intends to take two random samples of subjects from these populations and give to the two subject-samples the same sample of items, and test this hypothesis: $\mu'_y - \mu''_y = 0$.

It may be useful to use the high school remedial reading example here. The high school has 600 students who are eligible for the remedial reading program. Three hundred are randomly assigned to the special program, and 300 receive instruction under the "regular" program. It is important to note that the two groups of 300 subjects can be described in two ways. On the one hand, they can be viewed as two random samples from a larger population of 600 subjects. On the other hand, they can be viewed as two populations from which smaller samples can be randomly drawn. If this distinction is clear, the researcher's procedure is not too difficult. From the two populations of 300 each, he draws two random samples of subjects and gives the two samples of subjects one sample of items, perhaps 25 items, randomly sampled from the pool of 100 items. He uses the formulas suggested for Case II to estimate the mean and variance of the two populations of 300 subjects. Finally, a simple t-test, where

(12)

$$t = \frac{\bar{y}' - \bar{y}''}{\sqrt{\hat{\sigma}^2_{\bar{y}'} + \hat{\sigma}^2_{\bar{y}''}}} \text{ on } 2N - 2 \text{ df}$$

where

(13)

$$\hat{\sigma}^2_{\bar{y}} = \frac{\hat{\sigma}^2_{y}}{N},$$

would be appropriate to test the null hypothesis that the special remedial reading group is no different from the "regular" group. It should be recognized, however, that certain assumptions are being made by such a test. Specifically, we are, in effect, pretending that the items were not a sample implying that the y'_i and y''_i would be uncorrelated scores, (i.e., Cov $(\bar{y}', \bar{y}'') = 0$), in which case the usual t-test for independent samples could be used. In actuality, however, the y'_i and y''_i are not independent, being dependent on the particular item sample, which, in fact, has been drawn. The consequences of using the simple t-test are evident if we inspect $\sigma^2_{\bar{y}'-\bar{y}''}$, the sampling variance of the difference between the two sample means. As always,

(14)

$$\sigma^2_{\bar{y}'-\bar{y}''} = \sigma^2_{\bar{y}'} + \sigma^2_{\bar{y}''} - 2\, r_{\bar{y}'\,\bar{y}''}\sigma_{\bar{y}'}\sigma_{\bar{y}''}$$

where

(15)

$$r_{\bar{y}'\,\bar{y}''}\,\sigma_{\bar{y}'}\sigma_{\bar{y}''} = \text{Cov}\,(\bar{y}', \bar{y}'').$$

To the extent that the relative scores in the two experimental conditions are positively correlated (a condition which will be seen to most always prevail), $\sigma^2_{\bar{y}'-\bar{y}''}$ will be less than $\sigma^2_{\bar{y}'} + \sigma^2_{\bar{y}''}$ making the suggested t-test somewhat conservative. Again,

positive internal consistency is not difficult to meet in educational evaluation research. In other words, assuming nearly equal variances of and high positive correlations among item difficulties between the two groups, it will be advantageous to give fewer items to more subjects as long as the Kuder-Richardson internal consistency of the total item pool is positive.

CASE IV: Statistical-Psychometric Inference With Multiple Item Sampling and Multiple Subject Sampling

The methods described thus far use item sampling only superficially. A more efficient procedure is to administer different samples of items to different samples of subjects---that is, a random sample of subjects is drawn from the population of subjects and given a sample of items. Moreover, according to the formulas given in Case II, the mean and variance for the population of subjects taking the population of items is estimated. Another random sample of subjects can then be drawn non-overlapping with the first and given a sample of items. Estimates of the mean and variance are again made. This is repeated with non-overlapping, equal-sized samples of subjects. The estimates of the mean and variance are pooled (add up all the estimates and divide by the number of estimates) to provide a single estimate of the mean and a single estimate of the variance of the population of subjects taking the population of items. (See Appendix 1 for a computational example.)

In this use of item sampling, a number of non-overlapping samples of subjects are given a number of samples of items. A

number of questions immediately arise. For instance, how many people (or items) should be sampled? How large should the samples be? How should they be structured? In answering these questions some guidelines might be postulated. It is not necessary to sample the total population of subjects, although in many school settings it may be just as easy to do as not. It is important to sample the entire population of items, for the exclusion of merely a few items makes an important difference. It is important, also, to have every item responded to as often as every other item and, if necessary, data should be deleted to satisfy this suggestion. What's more, it is better to have every item paired with every other item at least once, and it is also desirable to have each possible pair of items responded to as often as any other pair. This last recommendation may not always be feasible. (Lord (1965) refers to Fisher and Yates (1938) tables 17-19 as providing assistance in fulfilling the recommendation.)

If it is possible to follow the guidelines, the procedure for fulfilling the guidelines is the optimal procedure for obtaining the most information given that number of observations. Of course, if resources permit, it is sometimes desirable to obtain more observations for particular research purposes. Following the guidelines also provides the best way of gathering more data.

CASE V: Statistical-Psychometric Inferences With Multiple Item Sampling and Multiple Subject Sampling in the Case of Two Subject Populations: Hypothesis Testing

Hopefully, the reader has noticed that Case III was little more than Case II, for it was extended to only two populations of

subjects, while Case IV was an extension of Case II to multiple item-subject samples. Case V is simply the next stage in this series of extensions. There are two populations and it is desired to use the procedures of Case IV for each of the populations and then to test a hypothesis about the means of populations. Stated in the framework of Case III and the remedial reading example, the situation is as follows: the high school has a pool of 600 students who are eligible for the special remedial reading program. Three hundred are randomly assigned to the special program; 300 are given the "regular" program. The mean and variance of the special group are estimated by the procedures of Case IV. The same is done for the "regular" group. Using these data on the two groups a t-test can be performed. A far more sensitive procedure*, however, would be to block on the multiple item samples in a two-factor analysis of variance design. Specifically, suppose 10 samples of items were given to 10 samples of examinees respectively in each instructional program. A 10 x 2 factorial design, then, can be used for the analysis, with the following advantages: (a) It will be possible to test not only for differences between instructional treatments, but for differences between item samples and the interaction between item samples and treatments. (b) The variance due to differences among item samples will be partitioned out of the error variance, making for a more sensitive test of treatment effects. (See Appendix 1 for an illustrative example of such an analysis.)

*Recommended by Lord in a personal communication.

There is an alternative procedure that can be used in this situation, and there are instances in which it may be more desirable. Consider the "regular" group of 300 subjects. Instead of using item sampling to estimate the mean and variance of the 300 subjects on the 100 items, it is suggested that just the mean of each item is estimated. This results in 100 numbers, each number an estimate of the mean of an item for the 300 subjects. The same procedure can be followed for the 300 subjects in the special reading group. Thus, 100 pairs of numbers are obtained, two estimated item means for each of the 100 items. A t-test for matched pairs can be performed on these data to examine the hypothesis of no difference between the groups. This test is quite powerful and also allows the researcher to report and examine the item data. Since the item data can be very useful for diagnostic purposes in the examination of the training program, this approach may be the most valuable one for many evaluation situations.

General Advantages of and Cautions About Item Sampling

The central idea of item sampling is simple. It is not necessary to give every item to every subject if one desires to estimate the performance of a group of subjects on a group of items. Much of this paper has been concerned with technical discussions of when item sampling would be an efficient data gathering procedure. In these technical presentations it is useful to examine the optimal circumstances for various uses of item sampling. These technical matters should not be allowed to obscure the fact that for many

educational issues, especially in the field of educational evaluation, item sampling provides the only viable method for collecting adequate data. The practical problems of educational research often prohibit the use of students for more than a short time but often do not seriously interfer with the testing of many students. In these frequent cases the issue is not whether item sampling will provide better parameter estimates than other procedures. The point is that item sampling can be used and that useful data can be collected. Perhaps, as mentioned on page 15, the t-test that can be used is quite cautious, perhaps overly so; but it can be performed. Item sampling not only permits, in many situations, more efficient data collection, but allows the educational researcher to perform some research which might not otherwise be possible.

However, one should not ignore two aspects of item sampling work. The first has already been mentioned: it is, that item sampling is set up to assist in the estimation of group statistics. The ordinary use of item sampling does not lead to statements about the performance of the individual subject. The other aspect of item sampling which should not be ignored is the fact that it is predicted on the assumption that the response of a subject to an item is independent of the context in which the item is presented. This assumption is made whenever one works with tests and test theory; but it is probably of more importance in item sampling, since the subject receives only a few items. Because of this, it is suggested that the researcher not conduct one item tests. What the size of the smallest test should be is not known, but tests of three to five items would seem to be as

small as is desirable.

Some Possible Uses of Item Sampling

To date, the authors have been able to find only one published article which actually used item sampling (Plumlee, 1964). It is obvious, therefore, that further empirical research is required to examine the actual performance of item sampling techniques in practice. These studies should be performed in situations where it is possible to collect data in several ways, and should compare various ways of estimating population parameters on the basis of partial data.

But there are several possible uses of item sampling which do not depend on the advantages of item sampling as opposed to other techniques of data collection. These are uses of item sampling where any other manner of collecting data is inappropriate. Two of these will be described briefly.

Most classroom tests are constructed for the purpose of differentiating among students. The best test is said to be the one which has maximum variance since that test will make it easiest to assign grades. This criterion tends to eliminate those items which every student either passes or misses. Items at about the 50 per cent difficulty level are said to be better than other items. In short, therefore, items constructed to ascertain whether course objectives have been achieved are often eliminated from classroom tests since they are often items which most students get correct; the good instructor achieves his goals and those items directly related to his goals are often passed by most students. Item sampling provides a technique for obtaining

Appendix 1
Computational Examples

Case I

a) Statistical inference: $N = 10$, $\bar{n} = 5$, and $x_{gi} = 1$ if item scored "correct" or $x_{gi} = 0$ if item scored incorrect.

		Person										$\sum_{i}^{N} x_{gi}$	p_g
		1	2	3	4	5	6	7	8	9	10		
	1	0	1	0	1	0	1	1	0	0	1	5	.5
	2	1	0	0	1	0	0	0	0	1	1	4	.4
Item	3	1	1	0	1	1	1	0	0	0	0	5	.5
	4	0	1	1	1	1	0	1	0	1	0	6	.6
	5	1	0	0	1	0	1	0	0	1	1	5	.5
$\sum_{g}^{\bar{n}} x_{gi}$		3	3	1	5	2	3	2	0	3	3		
y_i		.6	.6	.2	1.0	.4	.6	.4	0.0	.6	.6		

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{5.0}{10} = .5 = \hat{\mu}_y \quad \text{Calculated using (1)}$$

$$\hat{\sigma}_y^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N - 1} = \frac{.65}{9} = .07 \quad \text{Calculated using (3)}$$

b) Psychometric inference: $\bar{N} = 10$, $n = 5$. For purposes of example, simply replace N with $\bar{N}$ and $\bar{n}$ with n in above matrix.

$$\bar{p} = \frac{\sum_{g=1}^{n} p_g}{n} = \frac{2.5}{5} = .5 = \hat{\mu}_p \quad \text{Calculated using (5)}$$

$$\hat{\sigma}_p^2 = \frac{\sum_{g=1}^{n} (p_g - \bar{p})^2}{n-1} = \frac{.02}{4} = .005 \text{ Calculated using (7)}$$

Note: $\bar{y} = \bar{p} = \hat{\mu}_y = \hat{\mu}_p = .5$

Case II Statistical - psychometric inference

Suppose $\bar{N} = 20$, $\bar{n} = 10$ and we have reason to believe that σ_y^2 is approximately .1 and σ_p^2 is approximately .01. Furthermore, we are restricted to 50 item-subject observations for reasons of time and money. Since σ_y^2 is large relative to σ_p^2, it is advantageous to sample fewer items, say n = 5, and more people, say N = 10. For example purposes, suppose we get the same data matrix used in Case I.

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{5.0}{10} = .5 = \hat{\mu}_y \quad \text{Calculated (1)}$$

$$\bar{p} = \bar{y} = .5$$

$$s_y^2 = \frac{N-1}{N} \hat{\sigma}_y^2 = \frac{9}{10} (.07) = .065 \text{ Calculated using (3)}$$

$$s_p^2 = \frac{n-1}{n} \hat{\sigma}_p^2 = \frac{4}{5} (.005) = .004 \quad \text{Calculated using (7)}$$

$$\hat{\sigma}_y^2 = \frac{N(\bar{N}-1)}{\bar{n}(n-1)\bar{N}(N-1)} \left[n(n-1) s_y^2 - (\bar{n}-n) \left\{ \bar{y}(1-\bar{y}) - s_p^2 \right\} \right]$$

$$= \frac{(10)(19)}{(10)(4)(20)(9)} \left[5(9)(.065) - (5) \left\{ .5(.5) - .004 \right\} \right] = .05$$

Calculated using (10)

$$\hat{\sigma}_p^2 = \frac{n(\bar{n}-1)}{\bar{N}(n-1)(\bar{n})(n-1)}\left[\bar{N}(\bar{N}-1)s_p^2 - (\bar{N}-N)\left\{\bar{p}(1-\bar{p}) - s_y^2\right\}\right]$$

(Symmetrical analogue of (10))

$$= \frac{5(9)}{20(9)(10)(4)}\left[10(19)(.004) - (10)\left\{.5(.5) - .065\right\}\right] = -.007$$

Since σ^2 can not be negative, our best estimate of $\hat{\sigma}_p^2 = 0.000$.

Note: The difference in estimated values for σ_y^2 and σ_p^2 between the computational procedures of Case I and Case II is two hundredths and five thousandths respectively.

Case III: Statistical-psychometric inference with two subject samples and one item sample: simple hypothesis testing

Suppose we have access to 50 students and we want to compare programmed instruction A with programmed instruction B. Suppose we also have a fairly homogenious pool of 20 items related to the instructional content but are restricted, in terms of economy of time and money, to only 100 item-person observations. Since the test has a positive internal consistency, we can, for example, administer 5 items to each of 10 students in each instructional treatment. That is, we conceive of the 50 students as constituting two populations of 25 students each, from each of which we randomly select 10 students and assign one group to treatment A and the other to treatment B. 5 items are selected randomly from the pool of 20 and administered to each of the students after their instructional treatments.

We thus have two data matrices like that presented for Case II above. For purposes of example, suppose the data matrix for A is that of Case II and the statistics we need for treatment B have been

calculated. We then have the following data:

$N' = 10 \qquad N'' = 10$

$\bar{y}' = .5 \qquad \bar{y}'' = .9$

$s^2_{y'} = .065 \qquad s^2_{y''} = .050$

Following the recommendation made in the paper, we can treat the samples as being independent and use the usual t-test.

$\hat{\sigma}^2_{\bar{y}'} = .006, \hat{\sigma}^2_{\bar{y}''} = .004$ Calculated using (13)

$t = 4.00$ on 18 df calculated using (12)

$H_o: \mu_y' = \mu_y''$ is rejected at $p < .001$

Case IV: Statistical-Psychometric Inference With Multiple Item Sampling and Multiple Subject Sampling

Suppose we have a 50 item test and wish to get norms for this test on 200 people. We have, however, only the time or money to make half of the 10,000 item-person observations possible. One alternative is to use either Case Ia (administer, say, all 50 items to a sample of 100 people - 5000 observations) or Case Ib (administer, say, a sample of 25 items to all 200 people - 5000 observations). In each of these cases, either subject or item information is being lost.

Suppose, however, that we randomly select 5, non-overlapping, item samples of size 10 and administer them to 5, non-overlapping, random, examinee samples of sixe 40 respectively. This would amount to 400 observations per sampling or 2000 observations total.

Each of the 5 sampling procedures would yield a 10 x 40 Case II - type data matrix for which $\hat{\sigma}^2_y$ and $\hat{\mu}_y$ (and any other desired parameter estimate) could be computed as illustrated in

Case II. Suppose the following data was obtained for each of the five samples:

Sample	$\hat{\sigma}_y^2$	$\hat{\mu}_y$
1	.20	.63
2	.13	.41
3	.07	.55
4	.09	.82
5	.19	.57

Estimates for the entire item-subject population are obtained by taking simple arithmetic averages of the sample estimates. Hence,

$$\hat{\sigma}_y^2 = \frac{.20 + .13 + .07 + .09 + .19}{5} = \frac{.68}{5} = .14$$

$$\hat{\mu}_y = \frac{.63 + .41 + .55 + .82 + .57}{5} = \frac{2.98}{5} = .60$$

Case V: Statistical-Psychometric Inferences With Multiple Item Sampling and Multiple Subject Sampling in the Case of Two Subject Populations: Hypothesis Testing

Suppose the problem in Case IV is modified as follows: the 50 item test is an achievement test on the content covered by two alternative programmed instructional sequences A and B. We have a group of 400 students which we randomly assign to the instructional treatments. Our purpose is to test the effectiveness of treatment A relative to treatment B with the same economic restriction--that is, only 5000 total item-examinee observations are possible.

We can simply extend the procedures outlined in Case IV to our second sample of students. In other words, both examinee samples are randomly divided into 5, non-overlapping, samples of size 40 and given the same 5, non-overlapping, random samples of 10 items respectively. The amounts to a 5 x 2 analysis of variance design with 40 observations per cell which can be analyzed as follows:

Source	df
Item blocks	4
Treatments	1
Blocks X Treatments	4
Error	390

Appendix 2

Glossary of Terms

1. i: Lower case subscript used to identify the ith person.
2. g: Lower case subscript used to identify the gth item.
3. μ_y: Mean relative score on the population of items.
4. $\hat{\mu}_y$: Estimated mean relative score on the population of items.
5. μ_p: Mean item difficulty for the population of examinees.
6. $\hat{\mu}_p$: Estimated mean item difficulty for the population of examinees.
7. n: Number of items in a sample of items.
8. $\bar{n}$: Number of items in the population of items.
9. N: Number of people in a sample of people.
10. $\bar{N}$: Number of people in the population of people.
11. p_g: The difficulty of item g.
12. $\bar{p}$: Mean item difficulty for the given sample of people.
13. ρ: Population Pearson product-moment correlation coefficient.
14. r: Sample Pearson product-moment correlation coefficient.
15. s_p^2: Variance of the sample of item difficulties.
16. σ_p^2: Variance of the population of item difficulties.
17. $\hat{\sigma}_p^2$: Unbiased estimate of the population variance of item difficulties.
18. $\sigma_{\bar{p}}^2$: Sampling variance of mean item difficulty.
19. $\hat{\sigma}_{\bar{p}}^2$: Estimate of the sampling variance of mean item difficulty.
20. s_y^2: Variance of the sample of relative examinee scores.
21. σ_y^2: Variance of the population of relative examinee scores.

22. $\hat{\sigma}_y^2$: Unbiased estimate of the population variance of relative examinee scores.

23. $\sigma_{\bar{y}}^2$: Sampling variance of mean relative score.

24. $\hat{\sigma}_{\bar{y}}^2$: Estimate of the sampling variance of mean item difficulty.

25. x_{gi}: Score on item g by examinee i.

26. y_i: Proportion of items in test answered correctly by examinee i, i.e., relative observed score of examinee i.

27. $\bar{y}$: Mean relative score on the sample of items.

28. Cov (y', y"): Covariance between relative scores for the 2 experimental groups.

REFERENCES

Cook, D. L. and Stufflebeam, D. L. Estimating test norms from variable size item and examinee samples. Journal of Educational Measurement, 1967, 4 N.1, P. 27-33.

Fisher, R. A., and Yates, F. Statistical tables for biological, agricultural and medical research. Edinburgh: Oliver & Boyd, 1938.

Lord, F. M. Item sampling in test theory and in research design. Princeton, New Jersey: ETS, RB-65-22, 1965.

Plumlee, L. B. Estimating means and standard deviations from partial data - An empirical check on Lord's item sampling technique. Educational and Psychological Measurement, 1964, 24 (3), 623-630.