Studies of individual subjects: Logic and methods of analysis

Jum C. Nunnally* and **William E. Kotsch**

A statistical logic was developed by Payne & Jones (1957) and elaborated by Knight & Shelton (1983) for investigating the statistical significance or score changes for individual subjects. We decided that the particular approach advocated by Knight & Shelton would not be useful to researchers. These shortcomings are discussed and illustrated. More importantly, we took this as an occasion to discuss, in a general way, the available research designs and methods of analysis applicable to the study of individual subjects.

This article was inspired by a paper from Knight & Shelton (1983), which represented an extension of ideas presented earlier by Payne & Jones (1957). In the abstract to the paper by Knight & Shelton, it was said: 'The tables were prepared using Payne & Jones (1957) method of predicting the retest score by means of simple linear regression and testing the *significance* of the discrepancy between obtained and predicted retest scores' (italics added). It will be shown that the statistical methods proposed by Knight & Shelton have little to say about statistics for the investigation of individual cases or tests of significance. Apparently there are some fundamental misconceptions regarding the logic of investigating changes in test scores or other measurements for individual subjects. Our intention is to place in a wider perspective issues regarding both single-case research and the logic of studying human change.

The issue raised by Knight & Shelton is a familiar one in clinical psychological research and practice. An example would be that in which a mental patient is administered a standard test of intelligence before and after a study relating to the effectiveness of a new drug. It is observed that patient *A* changes from an IQ of 90 before the drug treatment to 120 after the drug treatment. Being statistically minded people, psychologists are prone to be sceptical of such observed changes and rightly want to document them in some statistical way. It might be asked in this case whether the observed change of 30 points in IQ could be due to 'chance', with respect to whatever meaning the word chance might have in this instance. However, at this point it is easy to make a premature flight into statistics without properly considering the logic and technique of investigating changes in individual subjects either naturally over time or as the result of an experimental treatment. In order properly to consider this matter, it will be necessary to digress into some familar issues which might not have been brought to bear on the study of human change.

Use of statistics in science

Inferential statistics concern the formulation of probability statements relating descriptive statistics found in samples of subjects to the descriptive statistics (called parameters) that would be obtained if a hypothetical population were investigated. The concept of sampling from a population (domain or universe) is an integral part of the formulation of inferential statistics.

As we view the matter, the most valuable products of inferential statistics are *confidence intervals* relating sample estimates to population parameters. Thus, employing various statistical models, one may formulate confidence intervals for nearly all of the major descriptive statistics that are employed customarily in research, including means,

* The Editor notes with regret that Jum C. Nunnally died on 21 August 1982.

0144-6657/83/020083–11 $02.00/0

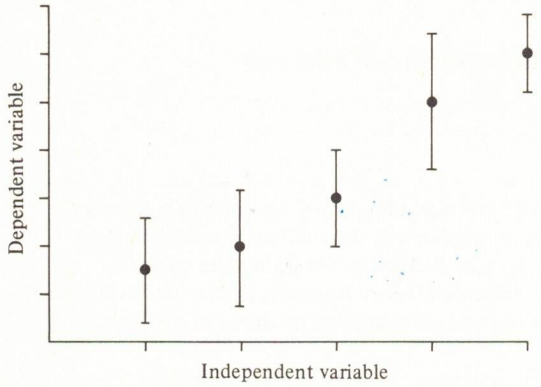

Figure 1. Relationship between an independent variable and dependent variable, with 95 per cent confidence intervals shown at each measurement occasion.

correlation coefficients, ratios of variances, multiple and partial correlations, canonical correlations, and others. An example of employing confidence intervals in an experiment is given in Fig. 1, which depicts the effects on a dependent variable of five treatment conditions. The five treatment conditions might be thought of as five levels of dosage of a drug or as five testing times in a developmental study of children. We feel it is very important in such investigations to calculate and depict confidence intervals about each point that is measured (see Morrison, 1976, for the mathematical rationale for obtaining such confidence intervals).

Unfortunately, psychologists seldom compute confidence intervals for descriptive statistics, but rather they tend to rely heavily on 'significance tests' (which surely are misleadingly named). These tests concern the 'null hypothesis', which might be thought of as a special case of employing confidence intervals. Numerous persons (including Nunnally, 1960) have bemoaned the overemphasis in psychology on the null hypothesis and the under-utilization of confidence intervals. This issue is varied because the articles by Knight & Shelton and Payne & Jones discuss significance tests only and do not mention confidence intervals.

Study of change

Starting in the mid-1950s, the development of proper methodologies for the study of human change became almost an obsession with numerous people in psychometrics, developmental psychology, and other areas. Some issues regarding simple difference scores for the study of change were raised earlier by Thorndike (1924). Lord (1956, 1958) and McNemar (1958) discussed issues relating to simple change or gain scores and recommended various mathematical formulae. Subsequently there was a series of arguments in print regarding the logic of studying change and the most appropriate methods of analysis, which culminated in a well-known book edited by Harris (1963). But that distinguished set of papers by no means calmed the waters, and lively controversy has continued to the present time. There was an exhaustive analysis of problems of analysing simple change or gain scores by Cronbach & Furby (1970), a follow-up paper by Furby (1973), and writings by Tucker *et al.* (1966), Namboodiri (1972), Marks & Martin (1973), Nunnally (1973, 1975, 1982*b*), Overall & Woodward (1975, 1976), Fleiss (1976), Hersen & Barlow (1976), Williams & Zimmerman (1977), Labouvie (1980), Zimmerman *et al.* (1981). This is by no means an exhaustive listing of the papers that have appeared on the logic and method of studying human change, and issues pertaining thereto appear in numerous

articles and books that are mainly devoted to some other topic, e.g. developmental studies. It was somewhat surprising to see no mention of this extensive literature in the paper by Knight & Shelton.

In the case posed by Knight & Shelton, patient *A* mentioned above would have two scores, X_1 and X_2. Later, consideration will be given to the ways that they recommend analysing such pairs of scores, but one can think of all approaches as being special cases of linear combinations. In the most general form, a linear combination would be:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots b_kX_k.$$

The above general formulation of linear combinations is, of course, omnipresent in psychological theory and in methods of analysis. It underlies the 'general linear model' on which all analysis of variance and more complex ANOVA methods are based. Such linear combinations underlie factor analysis, multiple correlations, and more complex correlational approaches. It is important to realize that such linear combinations condense the multivariate information in a set of X scores into one composite variable Y. Thus, for the sake of formulating both descriptive and inferential statistics, one can look at the characteristics of the Y distribution and consider only the X variables in relation to how they helped determine the eventual characteristics of Y. The reason this point is being emphasized is that, no matter how one investigates human change as a linear function of individual measurements, one eventually deals with a single distribution of Y scores. This is the case in the simplest instance in which 'before' scores are subtracted from 'after' scores; the differences score $Y = X_2 - X_1$.

The four major possibilities for investigating linear combinations of scores in studies of human change are shown in Fig. 2. Subsequently, it will be made clear where the proposal by Knight & Shelton fits into this scheme, but first we would like to discuss this classification scheme more generally as a framework for investigating single cases.

People sampling and psychometric sampling

In psychological research there are two important types of sampling: the sampling of subjects and the sampling of test items. Psychometric sampling concerns the representativeness of research materials with respect to the types of generalizations one wants to make about data. For traditional psychological tests (e.g. a measure of intelligence), this concerns the item coverage with respect to the construct underlying the test. Such considerations regarding representativeness of content are not unique to conventional psychological tests but rather pervade all types of psychological research. For example, there has been considerable interest in recent years in the extent to which word lists employed in studies of verbal learning are representative of the word categories about which generalizations are made (Santa *et al.*, 1979). Such concerns have led experimenters on verbal learning to modify their procedures of analysis to take account of what has been called 'stimulus variability', and there is some controversy as to how that should be done. While in experimental psychology issues relating to content representativeness have been more of a bother than anything else, similar issues with respect to measures of individual differences in abilities and personality characteristics are quite germane to the study of human change.

Issues regarding psychometric sampling concern measurement reliability and measurement error as have been formalized in the domain-sampling model for reliability (Nunnally, 1978). Essentially the model considers the item content for any test or measure as being randomly sampled from a prescribed domain. The content sampling error then relates to how well the items correlate with one another on average and how many items are sampled for a particular test. This simple model has been extended to take account of

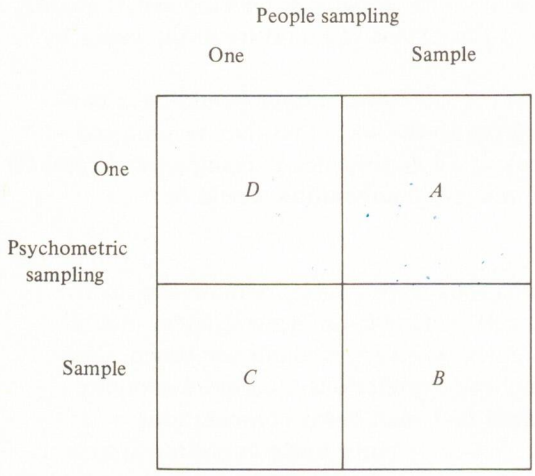

Figure 2. Four possible circumstances with respect to people sampling and psychometric sampling.

multiple factors contributing to the overall measurement error. Such multifactor models for reliability are said to concern generalizability (Cronbach *et al.*, 1972; Brennan, 1977; Brennan & Kane, 1979). It should be understood that what is meant by the sampling of content is sampling with respect to all of the factors that might be related to errors of measurement. Thus, psychometric sampling and psychometric error are based on the same notions of random sampling and probability as those for the sampling of people, although psychometric sampling is a much more complex matter.

Issues relating to psychometric and person sampling are depicted in Fig. 2. Thus, for either type of sampling, one may be concerned with inferential statistics relating to the study of one person in isolation from other subjects (e.g. the person who manifests a gain in IQ of 30 points), or one might be concerned with inferential statistics relating to groups of subjects (e.g. differences in mean performance of 30 subjects in each of three groups concerning the effectiveness of three drugs). Similarly, in psychometric sampling one may be concerned with only one measurement indicator in isolation from others, or one might be concerned with a group of such indicators. The issues here pertain not so much as to what information is available, but how one chooses to employ it with respect to inferential statistics. Thus, in the example given previously of the individual with the 30-point gain in IQ, there may have been numerous other subjects in the overall study, but one individual was singled out for some type of statistical scrutiny. With respect to psychometric sampling, there are usually multiple indicators; however, the data are often condensed into total scores, averaged scores, percentage correct scores, or in other ways using an overall descriptive statistic as a final measurement which does not explicitly take account of the underlying multiple indicators. Such mulitple indicators are the individual items on an IQ test or responses to individual items in studies of verbal learning, perception, and others. Some of the major issues concerning the employment of inferential statistics in the study of human change can be discussed with respect to the possibilities depicted in Fig. 2. Each of the four cells is discussed in turn.

Circumstance A: *People sampling and one measurement indicator*

There is little need to discuss in detail the situation depicted by possibility *A* given in Fig. 2, because this is the normal circumstance in which inferential statistics are employed. The usual practice is to have the individual appear only once in the data analysis in the form of

a total test score. Scores are combined across individuals and statistics are computed. The subjects, functioning as independent replicates, provide the necessary degrees of freedom for making inferences about parameters. The only major problem that is relevant here concerns the current controversy about the dangers of ignoring the 'lost degrees of freedom' arising from not considering the multiple indicators underlying global measures of traits. Not only is much information potentially thrown away by this procedure, but it is possible that the investigator is producing spurious findings by failing to take account of what actually happened with respect to the underlying item content (Santa *et al.*, 1979).

Circumstance B: *Both people sampling and psychometric sampling*

While it is possible to take account of the psychometric sampling underlying global measures when studying groups of people, the art of such statistical analyses is still in its infancy. There are some mathematical models that are intended to take account of both dimensions of sampling simultaneously, such as latent trait theory (e.g. Hambleton, 1980) and generalizability theory as referenced earlier. Unfortunately, such models are not equipped with the type of statistical distribution theory that would permit one to establish confidence intervals or other inferential statistics of a kind that are needed in the study of human change. Such models mainly produce interesting descriptive statistics concerning people's abilities and personality characteristics rather than inferential statistics.

A possibility seldom considered is to analyse each subject as though he or she were a separate experiment and then 'glue' subjects together in the context of the experimental design for groups of people. Thus, one could employ an ANOVA design with respect to the psychometric sampling underlying the responses of each subject and attach confidence intervals to linear contrasts in the data. The overall results can then be summarized in terms of grouping of subjects who have the same orderings of mean responses in intra-subject analyses. Overall probability statements can be made by converting probability statements made on individuals into normal deviates and averaging them. The necessary statistical methods have been discussed in recent literature pertaining to the aggregation of results from different experiments (Rosenthal & Rubin, 1979). While this is an interesting possibility the ideas are so new that little has been done with them.

Circumstance C: *One subject and psychometric sampling*

Thorough investigation of the 'single case' is possible only when there is acceptable psychometric sampling. Looking back at Fig. 2, the data could have come from a single subject at five points in time or in relation to five treatment conditions. The points on the curve could represent total scores on a test, average scores, or any other type of psychological measure. How would one obtain confidence intervals about those points? This is the prime question in the use of inferential statistics in studies of individuals. If there are multiple indicators underlying each point (e.g. 50 test items or 50 items in relation to learning or perception), then an attractive possibility is to employ powerful methods of statistics in relation to the multiple measurement indicators analogous to the way one would employ such statistics when the variability about points of central tendency are in relation to different persons rather than different items.

Employment of analysis of variance and other powerful methods of analysis within individual subjects requires an honest effort to sample content. However, the legitimacy of content sampling is not nearly as big a problem as is the requirement of experimental independence of data points that are counted as degrees of freedom. Whereas ANOVA and related statistics are surprisingly robust with respect to most underlying assumptions for the model, they are not robust when there are serial dependencies among responses from one subject. This problem would occur, for example, if a subject's response to the first item

on a test provided hints about how to respond correctly to the second and later items. Serial dependency would occur if the subject's failure to respond well to early items on a test discouraged him/her from responding correctly to later items. The problem of serial dependency is even more likely to occur in tests of personality, values, and interests. For example, there is a tendency in some tests of personality for the subject to 'commit' himself/herself with regard to different types of items on the test. If an individual rated himself/herself as being very friendly on one of the items early in the test, this commitment might influence subsequent ratings on other items related to sociability. When any form of serial dependence occurs among multiple responses from an individual subject, the precise number of degrees of freedom is in doubt and statements of probability cannot be trusted. However, very little research has been done to find out how much serial dependence actually occurs in typical tests of human abilities and personality and in the measures typically employed in psychological experiments. If such dependence is slight, the typical inferential statistics may be 'robust' with respect to this influence. Also, even if there is substantial serial dependence involved in sampling multiple responses from individual subjects, it may be possible to employ time-series analysis and related statistics (Box & Jenkins, 1970; Anderson, 1971; Glass *et al.*, 1975; Cook & Campbell, 1979). Both of these possibilities need to be explored extensively. If both prove to be unsuitable for handling potential serial dependence, then there is no real basis in inferential statistics for investigating single cases anywhere in psychology.

Even if it were not for the potential problem of serial dependence, one cannot employ inferential statistics with many types of tests and other measurements used in psychology, e.g. the Wechsler Scales, for three reasons. First, many tests are not homogeneous with respect to content but contain subscales. Secondly, most tests are 'composed' rather than sampled in any reasonable sense. Thirdly, for many tests such as the major individually administered tests of intelligence, all subjects are not administered all items.

There are situations in which it is reasonable to investigate single subjects with inferential statistics relating to psychometric sampling. This is the case with the computerized tests being developed by the authors (Nunnally, 1982*a*). A simple example is that of a test of numerical computation which consists of addition problems containing two rows of two digits. The computer composes such problems on a random basis, and consequently they are sampled. Since there are 10000 possible arithmetic problems of this kind, the domain can be considered infinite for all practical purposes. With this and other computerized tests, sets of responses are obtained on numerous occasions from individuals in studies of physical illness, drug effects, and experimental treatment conditions. While mindful of the possibilities of serial dependence as mentioned previously, we employ ANOVA and other powerful methods of inferential statistics to the experimental results from individual subjects. The results obtained from individuals can be woven together for the persons who manifest the same pattern of responses. Thus, some subjects may improve throughout a course of treatment for an illness and other subjects may decline in their abilities. The thorough investigation of single individuals allows us to document with inferential statistics what happens in each case.

Circumstance D: *One subject and no psychometric sampling*

Returning to Fig. 1, if one subject is studied and multiple indicators are not available for measurements, or if such multiple indicators are ignored and only a summary score is analysed, then there is no legitimate way to set up confidence intervals around the data points in the figure. Without psychometric sampling, each data point for a single subject cannot be documented with inferential statistics. Thus, there would be no way of saying

that the individual increased significantly from the first to second occasion and decreased significantly after that. Any attempt to study the single case without psychometric sampling at each point of observation is doomed to failure.

The situation described by Knight & Shelton, namely the individual who gained 30 points of IQ, is definitely in circumstance *D*. Here we are discussing one individual separately and would like to employ some inferential statistics relating to the size of the apparent change. Since the problem is discussed only with respect to composite scores (IQs), there is no opportunity to utilize inferential statistics based on psychometric sampling.

Knight & Shelton propose to get around this problem by borrowing a control group from another study and formulating a *post hoc*, quasi-experimental design. The experimental subject (the individual under study) and the control subjects receive a pre-test and a post-test. Between test the subject of interest was exposed to some kind of treatment. Did the treatment affect the dependence measure, the IQ? Before applying inferential statistics to help answer this question, we must consider the degrees of freedom available for a statistical test. There are adequate degrees of freedom in the pre-test group and the post-test control group. But there is only one subject in the post-test experimental group and no degrees of freedom, thus no statistical test involving the experimental group ($n = 1$) is possible.

In dealing with this problem, Knight & Shelton confuse descriptive statistics with inferential statistics. When information from a control group is utilized, statistics are generated. But, in regard to the individual of interest, these statistics are descriptive, not inferential. To illustrate this point let us take the case where no statistics are borrowed but rather all of them are generated in the context of an ongoing study. This would be the case if several hundred subjects were administered an intelligence test soon after they came to a mental hospital and again 3 months later. All of the necessary descriptive statistics could then be generated on this large sample of subjects. The end result of the statistical analysis would be a linear combination that was, let us say, approximately normally distributed. What could we say now about person *A* who showed the score change of 30 points? With regard to the mean and standard deviation in the study, it might be found that a change of 30 points was at the 99th percentile in terms of amount of change. Could one say that this result was 'significant at the 0·01 level?' Of course not. All that one could say was that subject *A* changed more than 99 per cent of the subjects. In this case, there would be no mechanism for establishing a confidence interval about the change of 30 points, or even discussing whether the apparent change was significantly different from no change at all. Indeed, it might be that individuals who changed by only five score points were manifesting non-chance changes. (It is surprising that anyone would want to employ such descriptive statistics as though they were inferential statistics because all that one could do would be to wait for the rather odd individual who was so normatively distinct from his/her group as to be regarded as 'significant' by this rather perverse statistical standard. For example, if one played this specious game and established in advance a significance level of 0·01, one would know before any research was undertaken that only one person in 100 would be found worthy of continued investigation).

It is more appropriate to say that such statistics concern degrees of unusualness rather that statistical significance in the proper sense. Discussions of statistical significance make sense when there are at least one, and usually numerous, 'degrees of freedom' underlying the descriptive statistic as when determining the statistical significance of differences between the means of two groups of subjects or when determining the statistical significance of a correlation coefficient. Because in the situation discussed by Knight &

Shelton there are no degrees of freedom available with respect to the first or second testing, there is no way to develop inferential statistics regarding the observed change of one individual.

Compromise solutions

Having reviewed the possibilities for psychometric and person sampling in relation to inferential statistics, what if anything can be said about the observed score change for person *A*? Although no one apparently has stated it as such, a compromise procedure had grown over the years in which group statistics concerning psychometric sampling (circumstance *B*) are borrowed for an analysis of score changes in individuals. The aim is to separate from the observed score the error due to psychometric sampling. The appropriate statistics for this task are reliability coefficients: measures based on internal consistency (coefficient alpha), correlations between alternative forms, and correlations obtained from retesting over various time intervals. If a full complement of such statistics is available from a large scale study conducted in standardizing the test, one can borrow such statistics for making probability statements about changes of scores in individuals. Such inferential statistics would relate to psychometric sampling error, as reflected in the various reliability estimates, rather than inferential statistics based on the sampling of people. With such borrowed statistics available, numerous formulae have been developed over the years, starting with the proposals by Lord (1956). Essentially, what is done in the equations developed from this line of reasoning is to postulate a *true change score* and then use all of the information in the first testing and the second testing (and in later testings if they are available) to estimate this hypothetical true change score.

This approach must be used with caution. One is never sure that the borrowed statistics are appropriate to the data being analysed. The reliability estimates obtained from previous research are just that–estimates, based on a model, interpretable only in the context of the kinds of generalizations one intends to make about data, and subject to error. One of the major problems with all such attempts to estimate true scores (true gain scores or any other types of true scores) is that it is necessary to regress scores toward the mean of some group. However, it is an open question to which group a subject 'belongs'. Without any other knowledge of the subject, the safest course is to compare the individual's score with the general norms developed for a test, e.g. a mean IQ of 100 and a standard deviation of 15 for the WAIS. However, that would not be wise if the subject were tested in a home for mental retardates or tested among the students at Oxford. In these cases, one could make a cogent argument that it would be more sensible to regress the individual's score toward the mean of the group from which he had been drawn. Even if there were no arguments regarding the mean of which group to employ in regressing scores, or which reliability estimates would be appropriate, one surely would wonder about the accuracy of all of the statistics that were borrowed. For example, many of the statistical estimates quoted by Knight & Shelton are based on so few subjects as to be untrustworthy.

For the foregoing reasons, it is apparent that, in most circumstances in which one would want to develop confidence intervals and tests of significance relating to score changes for individuals, only rough rules of thumb are available. A simple, sensible rule of thumb can be applied as follows. First, the investigator should regress on the first occasion of measurement toward the group mean by subtracting the mean from each score and then multiplying each such deviation score by the reliability coefficient. As mentioned previously, the group mean would be that of any group that has been designated for research, regardless of whether or not the group members as a whole are typical of any wider population of persons. If there is no obvious reference group for the individual, it makes sense to regress scores toward the mean of the sample on which the test was standardized, e.g. toward the mean of 100 on the WAIS. The regression coefficient in this case is the

internal consistency reliability of the test, as measured by coefficient alpha or the special version Kuder–Richardson Formula 20 used for dichotomous scores (Nunnally, 1978). Special formulae are required to estimate the internal consistency reliability when the test contains subtests, such as the WAIS (Cronbach *et al.*, 1972). If, rather than employing a retest on the second occasion, one employs alternative forms, then the regression coefficient would be the correlation between the two test forms administered on the same day, if such a statistic is available. One would not regress such scores in terms of the correlation of the first testing with the second testing (as Knight & Shelton propose), because that correlation mirrors not only the amount of reliable variance from occasion to occasion, but also the interaction of subjects with whatever treatments or natural sources of change occurred between the first and second testing.

If one is interested in absolute amounts of change rather than only relative amounts of change of subjects, one would then take the regressed scores on the first occasion and subtract them from the obtained (not regressed) scores on the second occasion. Of course, by this approach, the scores could all be positive or all negative if the means were substantially different on the first and second occasions. If it were known from previous studies that the means of retest scores were almost identical when no treatment condition intervened, no manipulation of mean differences would be needed. On the other hand, if it were known that a considerable practice effect was involved, the mean differences between the two testings could be adjusted to take account of this simply by subtracting that difference from each of the difference scores. Each difference score would then be divided by the standard error or measurement for the difference between two scores (Nunnally, 1978). If the obtained ratio was 2·0 in either the positive or negative direction, one could feel confident at approximately the 96 per cent level that a 'real' change had occurred. 'Real' change in this case is change attributable to influences not accounted for in the reliability statistic employed.

If one were interested only in relative changes, rather than absolute changes, then the mean of the difference scores would be subtracted from each difference score. The same standard error of measurement formula would then be applied.

The statistical rule of thumb is applied to person *A* as follows. Person *A* was a participant in an experiment concerning the effects of drug therapy on IQ and a number of measures of personality. Ninety subjects were assigned at random to three treatment groups. The IQ obtained for person *A* was 90; the mean IQ of the 90 subjects was 105. No information on reliability for the test was obtained from the study. A reliability estimate (coefficient alpha) of 0·80 was obtained in an extensive study done previously on the test. It is observed that the standard deviation reported in that study is very similar to that obtained for the 90 subjects in the experiment, so it seems reasonable to borrow the reliability of 0·80 for the current investigation. The score for person *A* would be regressed toward the mean score for his group by multiplying the deviate (−15) by 0·80, and this amount (−12) would then be added to the mean (105), providing a regressed IQ of 93. From previous studies of employing the test, there are apparently no observable practice effects when retesting is done at least two weeks later, and consequently the 'before' regressed score of 93 is subtracted from the 'after' score of 120, equalling a statistically adjusted change score of 27 (rather than 30). The standard deviations of scores on the first and second testings are almost identical and both are almost identical to the standard deviation reported in the previous research, namely 15·0. A correlation of 0·75 is found between the two testings. The squared standard error of measurement (SEM) for a variable (Nunnally, 1978) is obtained as follows:

$$(\text{SEM})^2 = b^2 x^2(1 - r_{xx}),$$

where b = a weight, x^2 = variance, r_{xx} = reliability. The equation for the SEM is more

typically shown when the weights for variables are 1·00 and thus do not appear. The squared SEM for a linear combination equals the sum of the squared SEMs for the separate variables. In the example used here, the first variable is weighted by the reliability coefficient, and thus the SEM squared is computed as follows:

$$(\text{SEM-1})^2 = (0{\cdot}80)^2\,(15)^2\,(1-0{\cdot}80) = 36.$$

To simplify the example, it will be assumed that the reliabilities at the two testing times are the same. The second variable is not weighted, and thus the squared SEM is obtained as follows:

$$(\text{SEM-2})^2 = (15)^2\,(1-0{\cdot}80) = 45.$$

The sum of the two squared SEMs is 81, and the squared root of that (9) is the SEM for the example. Since the score difference for person *A* of 27 is three times as large as the standard error of 9, one could have a high degree of assurance that the change is 'real' and not due to chance factors relating to measurement errors. One could go on to use the standard error of the difference scores to assert confidence intervals about the change of 27, but that would not be worth pursuing here.

It should be realized that the statistical procedures illustrated above are in relation to a particular model concerning measurement error, namely that in which measurement error is estimated with respect to internal consistency, with coefficient alpha and related estimates of reliability. The SEM obtained in this context does not take account of variations in the trait over time or other potential sources of measurement error such as that relating to test forms or test examiners. What may be assessed as error with respect to one model may be considered as 'real change' with respect to another model; this depends upon the types of generalizations made by the experimenter. One obvious alternative model would be to investigate changes as a function of some type of experimental treatment where a control group is available for comparison. For each group one can construct a regression equation in terms of the correlation between the first and second testings. Each data point in the treatment group can be compared with the regression line and standard error of estimate for the control group (Nunnally, 1975). Other refinements of the simple model illustrated above would be possible; however, these are seldom discussed in the psychometric literature and apparently almost never applied in practice.

Conclusions

It is obvious that one should avoid performing studies in relation to circumstance *D*, that in which only one subject is being considered and there is no recourse to psychometric sampling. The analysis given by Knight & Shelton is inappropriate. No matter how elaborately data from circumstance *D* are treated statistically, the best that one can employ are rough rules of thumb. Even those are brought into question when, as is frequently the case, there are opportunities to take advantage of chance in the investigation.

References

Anderson, T. W. (1971). *The Statistical Analysis of Time Series.* New York: Wiley.

Box, G. E. P. & Jenkins, G. M. (1970). *Time Series analysis: Forecasting and Control.* San Francisco, CA: Holden-Day.

Brennan, R. L. (1977). *Generalizability Analyses: Principles and Procedures.* Iowa City, Iowa: The American College Testing Program, Technical Bulletin, 26.

Brennan, R. L. & Kane, M. T. (1979). Generalizability theory: A review. In R. E. Traub (ed.), *New Directions for Testing and Measurement: Methodological Developments.* San Francisco: Jossey-Bass.

Cook, D. C. & Campbell, D. T. (1979). *Quasi-experimentation.* Skokie, IL: Rand McNally.

Cronbach, L. J. & Furby, L. (1970). How we should measure 'change'–Or should we? *Psychological Bulletin*, **74**, 68–80.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements*. New York: Wiley.
Fleiss, J. L. (1976). Comment on Overall and Woodward's asserted paradox concerning the measurement of change. *Psychological Bulletin*, **83**, 774–775.
Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology*, **8**, 172–179.
Glass, G. V., Willson, V. L. & Gottman, J. M. (1975). *Design and Analysis of Time Series Experiments*. Boulder, CO: Colorado Associated University Press.
Hambleton, R. K. (1980). Latent ability scales: Interpretations and uses. In S. Mayo (ed.), *Interpreting Test Performance*. San Francisco: Jossey-Bass.
Harris, C. W. (ed.) (1963). *Problems in Measuring Changes*. Madison, WI: University of Wisconsin Press.
Hersen, M. & Barlow, D. H. (1976). *Single Case Experiment Designs*. New York: Pergamon Press.
Knight, R. G. & Shelton, E. J. (1983). Tables for evaluating predicted retest changes in Wechsler Adult Intelligence Scale scores. *The British Journal of Clinical Psychology*, **22**, 77–81.
Labouvie, E. W. (1980). Measurement of individual differences in intra-individual changes. *Psychological Bulletin*, **88**, 54–59.
Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement*, **16**, 421–437.
Lord, F. M. (1958). Further problems in the measurement of growth. *Educational and Psychological Measurement*, **18**, 437–454.
Marks, E. & Martin, C. G. (1973). Further comments relating to the measurement of change. *American Educational Research Journal*, **10**, 179–191.
Morrison, D. F. (1976). *Multivariate Statistical Methods*, 2nd ed. New York: McGraw-Hill.
McNemar, Q. (1958). On growth measurement. *Educational and Psychological Measurement*, **18**, 47–55.
Namboodiri, N. K. (1972). Experimental designs in which each subject is used repeatedly. *Psychological Bulletin*, **77**, 54–64.
Nunnally, J. C. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, **20**, 641–650.
Nunnally, J. C. (1973). Research strategies and measurement methods for investigating human development. In J. R. Nesselroade & H. W. Reese (eds), *Life-span Development Psychology*. New York: Academic Press.
Nunnally, J. C. (1975). The study of change in evaluation research. In E. L. Struening & M. Guttentag (eds), *Handbook of Evaluation Research*, vol. 1. Beverly Hills, CA: Sage Publications.
Nunnally, J. C. (1978). *Psychometric Theory*, 2nd ed. New York: McGraw-Hill.
Nunnally, J. C. (1982*a*) Computerized testing of human abilities. In B. B. Wolman (ed.), *International Encyclopedia of Psychiatry, Psychology, Psychoanalysis, and Neurology*. New York: Van Nostrand Reinhold.
Nunnally, J. C. (1982*b*). The study of change: Measurement, research strategies, and methods of analysis. In B. B. Wolman (ed.), *Handbook of Development Psychology*. New York: Prentice Hall.
Overall, J. E. & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, **82**, 85–86.
Overall, J. E. & Woodward, J. A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin*, **83**, 776–777.
Payne, R. W. & Jones, H. G. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology*, **13**, 115–121.
Rosenthal, R. & Rubin, D. B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin*, **86**, 1165–1168.
Santa, J. L., Miller, J. J. & Shaw, M. L. (1979). Using quasi *F* to prevent alpha inflation due to stimulus variation. *Psychological Bulletin*, **86**, 37–46.
Thorndike, E. L. (1924). The influence of chance imperfections of measures upon the relation of initial scores to gain or loss. *Journal of Experimental Psychology*, **7**, 225–232.
Tucker, L. R., Damarin, R. & Messick, S. A. (1966). A basefree measure of change. *Psychometrika*, **31**, 457–473.
Williams, R. H. & Zimmerman, D. W. (1977). The reliability of difference scores when errors are correlated. *Educational and Psychological Measurement*, **37**, 679–689.
Zimmerman, D. W., Brotohusodo, T. L. & Williams, R. H. (1981). The reliability of sums and differences of test scores: Some new results and anomalies. *Journal of Experimental Education*, **49**, 177–186.

Received 1 April 1982; revised version received 2 August 1982

Requests for reprints should be addressed to William E. Kotsch, Department of Psychology, Vanderbilt University, Nashville, TN 37240, USA.
Jum C. Nunnally was formerly also at the above address.