JOURNAL OF EDUCATIONAL MEASUREMENT
VOLUME 20, NO. 4 WINTER 1983

MEASURING CHANGES IN EDUCATIONAL ATTAINMENT OVER TIME: PROBLEMS AND POSSIBILITIES

HARVEY GOLDSTEIN
Institute of Education, University of London

To attempt to measure trends over time seems a perfectly natural thing to do. It is a common activity, for example, when studying mortality rates, heights of children, or economic indicators such as the retail price index. In education the popular notion of "standards" is often linked closely to questions about changes in attainment or achievement over time. The Assessment of Performance Unit (APU) in Britain and the National Assessment of Educational Performance (NAEP) in the United States have both given a high priority to making inferences about trends in attainment over time (Gipps & Goldstein, 1983; Wirtz & Lapointe, 1982).

Despite much interest in the issue, there has been curiously little attempt to define precisely what is meant by "trends over time," or to discuss the attendant measurement problems. First, in order to set the scene, the ways in which the issue is perceived in areas other than education are briefly reviewed. Next a discussion of the logical basis of the most commonly used methods underpinning analyses of trends over time is provided. Finally, alternative formulations of the issue are presented, together with suggestions for their implementation.

OTHER AREAS

The measurement of mortality rate, children's height, and the retail price index will serve to introduce the general difficulties that arise when dealing with trends over time.

Table 1 (Goldstein, 1979) shows the perinatal mortality rates per 1,000 births for Austria and England and Wales in 1954 and 1965. Austria has a higher rate than England and Wales in both years, and for both countries, the rate is lower in 1965 than in 1954. Also in 1954 the difference between the rates is larger than it is in 1965.

Suppose, instead, that we look at the log (mortality rates), that is, we measure on a scale of ratios of rates (see Table 1). Using this scale, the difference between the countries is smaller in 1954 than in 1965.

Which scale should be used to make comparative statements, given that these are of importance in attempting to assess the relative merits of perinatal health services? In general there seems to be no clear answer; it will depend on the precise purpose of the measurement and the nature of the factors known to affect the mortality rates. For example, if linear models are used to study the effects of such factors, a choice between scales might be governed by which model required fewer parameters to give an acceptable fit. Thus, if a unit change in any of the independent variables in such a model produced a corresponding proportionate rather than additive change in the mortality rate, then a logarithmic scale would be the more appropriate. But when we have merely a simple comparison of rates without knowledge of the properties being measured, it is difficult to see how a satisfactory choice is to be made.

An earlier version of this paper was delivered at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, New York, March 19–23, 1982.

Table 1

Perinatal Deaths per 1000 Total Births

Country	Year	
	1954	1965
Austria	41.5	29.5
England and Wales	38.0	26.9
Difference	3.5	2.6

Turning next to trends in height, Table 3 shows that the mean differences between British boys and girls at age 4.0 and age 8.0 years are the same (Goldstein, 1979). Table 4, however, shows that when allowance is made for the increasing variability of height at age 8.0 by dividing all measurements by the appropriate standard deviations, then this standardized difference is smaller at age 8.0 than at age 4.0. Again, what scale should be used is unclear. Most human biologists would probably choose to use height itself, whereas those concerned with mental measurement might choose to standardize the distributions at each age.

Fresh problems arise in the case of the retail price index. In addition to the comparative problems encountered above, there is also the difficulty of defining a constant measuring instrument. Because the retail price index is used as a key measure of inflation, and indeed it is of little use unless it can provide a measure of change over time, great store is set by the movement of the retail price index, at least in the United Kingdom. Essentially, the index is a weighted average of different retail commodities, the weights being determined by typical consumer spending patterns. That, however, is the difficulty. Spending patterns change over time so that the index, if it is to remain useful, must also change correspondingly over time. At each change, the new index will be adjusted to the old one and then used until it, too, is updated. The problem, of course, is that there is no way of knowing whether the new instrument is really measuring the same thing as the old one. It is simply taken as an assumption, and to the extent that the assumption is shared by a large number of people, the retail price index could be said to comprise a useful measurement. This third example comes rather close to the case of educational achievement tests and, indeed, the analogy between an educational test as a shopping basket containing a weighted average of different individual items and the retail price index seems a good one.

Table 2

Log_{10} Perinatal Deaths per 1000 Total Births

Country	Year	
	1954	1965
Austria	1.618	1.470
England and Wales	1.580	1.430
Difference	0.038	0.040

Table 3

Mean Difference in Height (cm) of Boys and Girls

	Age	
	4.0	8.0
Boys - Girls	1.2	1.2
S.D. of Height	4.34	5.77

TRENDS IN READING STANDARDS

The above issues, which have been introduced through areas of knowledge often regarded as more secure than is education, can also be considered in relation to educational tests themselves. A useful illustration of the problem is found in the history of a series of national reading tests in England and Wales from 1948 to 1970, described by Start and Wells (1972). Two tests were given to 11- and 15-year-olds, the Watts-Vernon between 1948 and 1970 and the NS6 between 1955 and 1970. Figure 1 summarizes the results.

The crude picture presented in Figure 1 was used by many to conclude that reading standards had begun to decline during the 1960s, or at least had stopped rising, and that the blame for this resided in the schools. A government committee to look into the teaching and learning of English, the Bullock Committee, resulted partly from such concerns as did the so-called "Great Debate" about the quality of British education inaugurated by Prime Minister Callaghan in 1976. In addition, the APU also resulted partly from these findings (Gipps & Goldstein, 1983). Clearly the test results were important, but it was not long before critics began to question their relevance. For example, Burke and Lewis (1975) pointed out that the Watts-Vernon test was developed in the 1940s and contained words like "mannequin parade" which might have been familiar to 11-year-olds then but were not so familiar in the 1960s. To put it another way, such items had become biased against the students in that period. Thus, any falling off in the test scores might just as easily be attributed to an increasing test difficulty as to a lowering of achievement in the population.

Most tests of any value will eventually become outdated, and at least some of their items will require replacement. Two immediate questions are how long we may go on using a test before having to change it, and when it is changed, how we may continue to make statements about changes over time. In order to answer these questions, a general framework for using tests over time is developed in the next section.

Table 4

Mean Difference in Height (standardized) of Boys and Girls

	Age	
	4.0	8.0
Boys - Girls	0.28	0.21

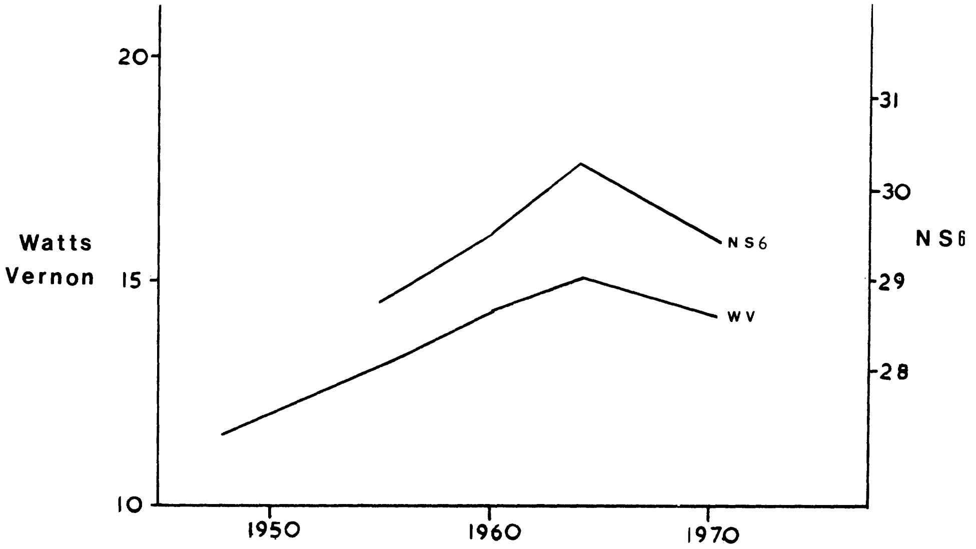

Figure 1. Mean test scores of 11-year-olds over time

TESTS OVER TIME

Consider three sets of items A, B, and C, all concerned with measuring a particular area of attainment for children of a specified age group, for example, arithmetic among 10-year-olds. Sets A and B of these items are for use at time one and set C for use at time two. We want to say something about the change in mean score between time one and time two using the items.

Following the earlier discussion, suppose that the set of items A are those judged relevant or appropriate at time one but not time two. For example, if they are concerned with language, then they may deal with language that has fallen out of use; if they are concerned with mental arithmetic, then they may have become uninteresting because the advent of cheap electronic technology between time one and two no longer requires children to be able to perform the operations used. In both examples, the items might supposed to have become more difficult because children would be less exposed to what they represent.

Suppose also that the reverse argument applies to the set of items C. These are relevant to time two but not to time one, again because of changing curricula or social uses. Finally, the set of items B are those which are relevant to both times.

While different people would undoubtedly form different views about which items fell into which category, the important thing to note about these categories is that it is only possible to define them with hindsight at time two. When we start with item sets A and B at time one, which items belong to A and which to B cannot be determined. Of course, we can forecast the future and attempt to design a set of B type items, but we still have the dilemma of not knowing whether our forecast is right until time two.

There are two inferences to be drawn from this argument. First, if we wish to use an unchanging or equally relevant set of items to measure change, then we cannot know what these items are until time two. In particular, we may find that set B actually contains very few items at all. Second, we have little control over which items fall into set B. Thus, those items that do end up in B may or may not be of interest, but what is fairly certain is that, out of all the items A, B, and C, they will represent only a partial aspect of what it is that we wish to measure at either time. In other words, those items which happen to remain equally relevant may well be a limited subset of the items of real interest, and because there is no certain way of knowing at the outset which items will be of real interest, this poses considerable problems for test construction.

There are three responses to the dilemma of changing content interests. The first simply is to ignore it: to declare a set of items to be the area of interest, as it were by fiat, and then proceed to measure all of them over time. Such an approach is possible, but presumably difficult to maintain as the credibility of individual items comes into question. This is effectively what happened to the National Foundation for Educational Research reading tests when they were criticized in the 1970s.

The second response, more subtle and reminiscent of the retail price index procedures, underpins the use of latent trait models, especially Rasch, for measuring trends over time. It runs as follows. Even if items in set A change their properties over time, so long as set B is not empty, we can in principle link the items in set A and likewise those in set C to those in set B (retrospectively, of course). If we use Rasch, then the difficulties of the A and C items are separately calibrated against the constant item B difficulties. In this case, however, we immediately confront a logical contradiction. If we consider the two times as part of a single time period, then the Rasch model, or any other model that assigns parameter values to items, requires that the characteristics of all the items during that period remain constant, otherwise the assumptions of the model are not met. Yet by hypothesis, the A and C items, but not the B items, change their characteristics or parameter values and, hence for these items, there can be no unique calibration and the model cannot therefore be true, over the whole time period.

An argument similar to the above can also be made for methods such as generalizability theory, where the reasoning applies to attempts to change the universe of items between times by deleting and adding appropriate items. The argument can also be extended to the use of subjective assessment by raters or examiners who, retrospectively, assess test responses from two times. Here the criteria used for assessment for some items (set B) will be equally relevant for both times, but for others (sets A and C) will be more relevant at one time or another.

Equating procedures based on total test scores also face the same problems because they contain a contradiction between tests changing in relative difficulty over time and the need to assume a constant equating function. The literature on test equating is curiously reticent on the question of specifying the reference population to which any particular function applies. Typically, there is an implicit assumption that single functions can be found which will apply to all population groups. Even for a single point in time, such an assumption is highly questionable, yet empirical evaluations of it seem to be rare (Goldstein, 1982). Thus, once it is asserted that educationally significant parameters change their values, then scale constancy does not exist and absolute comparisons over time cannot be made.

The third response is to accept the above difficulties and to ask the following question. If we start out with a test containing items A and B, for how long can we continue to use it so that we can reasonably be sure that it remains relevant? In the United States, for example, the NAEP has made the same set of comparisons of language items over a 10-year-period, and an Ontario test (Ontario Mathematics Achievement Test) has been used to make comparisons in mathematics over a 12-year-period. It is a problem, at least in theory, which confronts every test publisher.

There is a clear duality between attributing change in an item parameter value to a change in the population response or in the characteristics of the item. The point, however, is whether the item should be regarded as an equally fair assessment of the education system at each time, and it is here that judgment as well as empirical evidence is needed. In any assessment of this kind using constant tests, at the very least there should be an expert judgment of the results in light of such equity considerations.

With regard to the set of items B, there is the problem of how to extract the maximum useful information from a possibly very limited subset. Of course, there is useful information to be obtained from the very process of deciding which items do not belong to set B. The procedure for making such a decision should ideally involve consideration of changing curriculum content, social demands for knowledge and skills, and so forth; in other words, a considerable amount of careful research of interest in its own right. Moreover, it is possible to obtain at both testing times information, however crude, about item relevance for each child in terms of curriculum exposure or opportunity to acquire appropriate knowledge. This would provide further useful information and also serve as a check that an item truly does belong to set B. Finally, despite what has been said about set B possibly being very restricted, it does contain some information and, at least for that set of items, useful statements about absolute change over time may be possible. Nevertheless, it is bound to remain only a partial answer to the larger problem.

REFORMULATING THE QUESTION

Given the difficulties with absolute comparisons over time, two alternative approaches remain which may tell us something about change over time and actually provide more interesting information than absolute comparisons. First, even when it is not possible to make absolute comparisons, relative comparisons may still be possible. Given the incomparability of mean test scores, suppose we standardize typically different tests at each time so that they have exactly the same distribution over their respective populations and, in particular, the same mean value.[1] We may now study the differences between subgroups of the population, for example, regions, to see how these change over time. Thus we may find that the (standardized) difference between the north and south is less at time two than at time one. This would mean that, relative to the total variation in the population at each occasion, the regional differences have decreased. We can extend such comparisons, for example, to studying the variance between classrooms or schools, looking say at whether the percentage of explained variance has changed.

There are other ways of separately standardizing the distribution. All, however, would for simplicity take as their starting point an equating of average values (say the mean or median). It would also be convenient, although not necessarily always appropriate, to normalize the separate distributions, so that only one further decision, about the relative standard deviations, has to be made. We do not need to make them equal, but could instead, for example, equate measured group differences. Thus, say for reading, we could

[1] If, for convenience, we care to transform to normality, then we need only equate the means and standard deviations. There could be a debate about which standard deviations to use, for example, total or within-groups, but that is a side issue.

choose to make the average change in test scores over a 1-year-period (around the mean age of testing) equivalent at the two occasions; this would lead to relative comparisons in terms of reading age. This particular method of standardization need not be linear if the score to age conversion is nonlinear. In fact the use of reading age commends itself because it has a ready interpretation, and such interpretation possibilities should be closely linked to a choice of standardization procedure.

In general, different procedures will lead to different estimates for relative change. It is in this that the main difficulty lies. Unless a specific standardization procedure can be justified both on substantive grounds and ease of interpretation, or unless a wide variety of procedures produces similar empirical results, the choice will be arbitrary. It is obvious, therefore, that careful consideration needs to be given to an adequate justification of the choice of standardization procedure.

LONGITUDINAL MODELS

There is a parallel situation to the present one in the study of change in educational test scores with age, where again absolute comparisons over ages are ruled out because different measuring instruments are used. When different samples of children at different ages are compared, the above problems of standardization also arise. However, when the same children are measured at different ages, as in a longitudinal study, most of the problems disappear. A detailed discussion can be found in Goldstein (1979), but briefly what is done is to make comparisons between subgroups at time two by comparing the mean scores of those individuals in each group who have the same score at time one. This comparison is made for each score at time one and an overall picture, as in Figure 2 (from Goldstein, 1979, p. 124), is built up. The inference from this conditional relationship is that individuals who start from the same attainment at time one have different mean attainments at time two determined by the social group and, in that sense, have made different rates of progress. This approach has the property that scaling is irrelevant. All linear transformations leave inferences unaffected, as do nonlinear transformations for the time one scale.

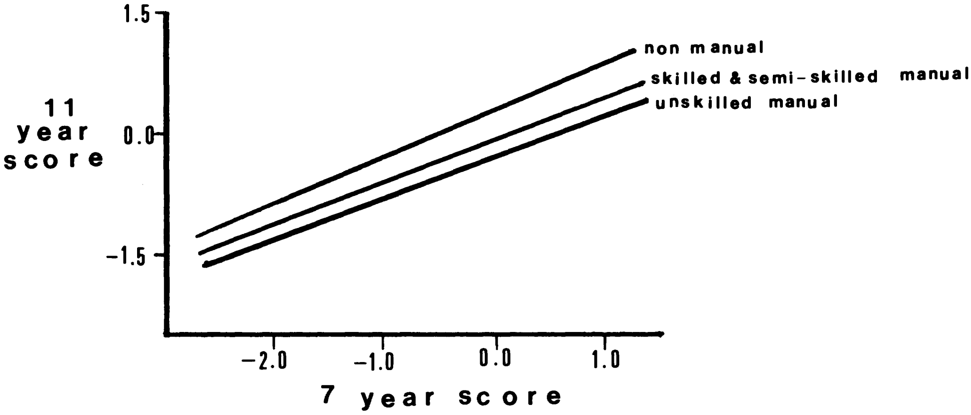

Figure 2. Fitted regression lines of 11-year reading score on 7-year reading score for three 7-year social class groups (standard deviation units; National Child Development Study data)

The conditional model motivates the second approach to studying relative change over time. In the example of regional differences, we would study the average score of the schools in each region at time two, for each given time one score. In the simplest case, we would fit parallel linear regressions of time two score on time one score, with the school as the unit of analysis. We would then be able to draw valid inferences about the relative change over time in mean school scores for the two regions. In much assessment work, the school seems to be the smallest natural unit that generally remains intact over time, although units at higher levels of aggregation, such as districts or education authorities, are possible. The interpretation of such analyses has to be considered carefully, but it seems a promising line to pursue.[2]

INFERENCES AND IMPLICATIONS

Even if we could meaningfully measure absolute trends in achievement over time, it would tell us very little. The lack of continued upward progress in the postwar series of reading tests in England and Wales, mentioned earlier, alarmed educationalists. Yet why should this have been so? There is no principle that states that an approximately linear trend will continue unless acted upon by some external force; indeed, there are myriad examples to the contrary. Furthermore, even when we resist such crude determinism, there is still the difficulty of explaining any particular pattern of change over time. If test scores decrease, why is this necessarily to be attributed to teachers or schools? It might just as well be a consequence of increasing amounts of pollution from vehicle exhausts. If we are not careful, we can all too easily succumb to the parallel time series fallacy, as exemplified in the legendary tale about the number of storks and the changes in the Swedish birth rate.

Thus, not only are there serious technical difficulties associated with measuring absolute trends over time, but there also seems to be little point in doing so. This point is perhaps the most important one given the determination of so many of those associated with education to try to say something about absolute changes over time. In the United Kingdom at least, an enormous effort, costing millions of pounds, has gone into a national monitoring program, APU, whose underlying motivation was to provide absolute comparisons over time. Indeed, what is so interesting about this enterprise is that it began to take shape in the early 1970s when there was growing interest in the accountability of the education system. The reading test results were used by antiprogressives to blame modern teaching methods for a decline in standards, and yet quite apart from the difficulties of item relevance, the implied causal link between achievement test changes and curriculum policies seems largely to have gone unquestioned. It was also at this time that caveats concerning item relevance and appropriateness were creeping in, thus throwing doubt upon the whole monitoring exercise, and the Rasch model appeared, promising to make absolute comparisons possible once more. Perhaps the single most

[2] There are some statistical difficulties with this approach, such as the exact form of regression relationship, the existence of measurement errors and so forth, but these do have solutions, at least in principle. The more difficult problems concern working with relatively unfamiliar units. In education, we are used to treating the individual child as the basic unit, and we make interpretations about factors that affect individuals. There is now, however, an increasing concern with the study of units at higher levels of aggregation, such as the classroom and the school, and the interactions between them (see Roberts & Burstein, 1980).

powerful reason for the perseverance of the Rasch model within the APU was precisely this possibility of measuring absolute trends over time (Gipps & Goldstein, 1983).

Measuring relative changes over time, using either standardized differences or longitudinal analyses, is both a technically feasible and substantively interesting exercise. If we discover that regional differences have narrowed and that this continues to remain the case even after a number of possible confounding variables have been allowed for, then we may have begun to uncover something interesting and useful. In effect this is merely another way of saying that even though we are limited typically to carrying out observational studies, the techniques of comparative analysis are available to investigate possible causal hypotheses, just as in other areas of social science.

In advocating such comparative analysis, the problems that remain to be solved are not being minimized (e.g., the choice of an appropriate standardization or unit of analysis model). Rather the intent is to direct attention toward a more fruitful area of interest than has been evident in past discussions of trends over time. If we ask the right kinds of questions, then we may begin to get some worthwhile answers.

REFERENCES

BURKE, E., & LEWIS, D. G. Standards of reading: A critical review of some recent studies. *Educational Research,* 1975, **17,** 163–174.

GIPPS, C., & GOLDSTEIN, H. *Monitoring children: An evaluation of the Assessment of Performance Unit.* London: Heinemann, 1983.

GOLDSTEIN, H. *The design and analysis of longitudinal studies.* London: Academic Press, 1979.

GOLDSTEIN, H. Models for equating test scores and for studying the comparability of public examinations. *Educational Analysis,* 1982, **4,** 107–118.

ROBERTS, K. H., & BURSTEIN, L., (Eds.). *Issues in aggregation. New directions for methodology of social and behavioral science (No. 6).* San Francisco: Jossey-Bass, 1980.

START, K. B., & WELLS, B. K. *The trend of reading standards.* Slough: NFER, 1972.

WIRTZ, W., & LAPOINTE, A. *Measuring the quality of education. A report on assessing educational progress.* Washington, D.C.: Wirtz & Lapointe, 1982.

AUTHOR

HARVEY GOLDSTEIN. *Address:* Institute of Education, 20 Bedford Way, London WCIH OAL, England. *Title:* Professor of Statistical Methods. *Degrees:* B.Sc., University of Manchester, England. *Specializations:* Educational measurement, longitudinal studies.