ON THE ACCURACY OF INDEX NUMBERS

BY BENT HANSEN
Department of Economics, University of California, Berkeley

AND EDWARD F. LUCAS
Institute of International Studies, University of California, Berkeley

Index number accuracy is affected by formula specification and sampling error. The authors argue that an index formula should be "ideal" and "exact" (with reference to the range of economically plausible aggregator functions) to be economically justified. These indices are invariant in the homothetic case, as well as in certain non-homothetic scenarios. Empirically, based on foreign trade data for Egypt from 1885–1961, the set of economically justified indices are virtually identical, supporting the theoretical argument that "instrumental error" or "formula variance" should be a negligible factor contributing to index number error. In a discussion of sampling error, on the other hand, the authors criticize earlier work and propose an upper and lower bound. Using the same data, these limits imply that sampling error may be a serious problem for many indices.

INTRODUCTION

Several years ago we began an empirical study of Egyptian foreign trade (Hansen and Lucas, 1978) based on comprehensive import and export indices for the period since 1885.[1] Initially we did not see the computation of the indices as a major economic problem, and were prepared to focus mostly on the results of our computations rather than the computations themselves. As we became more involved in the project, our dissatisfaction grew and solidified around two primary concerns: (1) a lack of guidelines for index formula specification, and (2) no accepted measure or even clear idea relating to index confidence. In pursuit of these concerns we encountered some of the incomprehension that separates economic theory from practical empiricism, as well as the fatalistic attitude that pervades discussions generally of indexing problems. In surveying the state of the indexing art in 1974, Samuelson and Swamy concluded on this point: "We must accept the sad facts of life" (p. 592).

While perhaps inevitable then, our survey of the contemporary situation is rather more optimistic. Recent theoretical work offers a considerable commentary on formula specification, which complements rather than contradicts the original work of the empiricists of the early part of this century. Moreover, there also exists a backlog in statistical theory which can be applied to our related concern involving index confidence. Therefore, in applying what is known but not applied to a variety of empirical situations, it should be possible to significantly improve the quality of work in the many economic applications dependent on index computations.

There also exists a darker side to our concern, which, however, is not dealt with in this paper. Our treatment of data assumes a level of idealization that in

[1]This research was supported by National Science Foundation Grant SOCSES80-06214 and Committee on Research, U.C. at Berkeley.

some ultimate reality doesn't exist, but which is invariably assumed in theoretical and even most empirical work. In preparing the price and quantity series which are the inputs to an index number calculation, the questions of "specified" prices or unit values, commodity definition and the level of aggregation, technological change, new products and trending quality change are important problems which could have a preponderant effect on the final result, as they could in a regression or any other calculation. On a philosophical level it has to do with the correspondence between theoretical concept and measurable phenomenon and is a general economic problem rather than being peculiar to index numbers.

I. Index Formula Specification

A. *Analogue Tests*

In general, the purpose of index numbers is to enable the quantitative treatment of useful composite commodities:

> "In a loose qualitative description, such terms as "real wage" and "producer goods" may simply indicate the totality of commodities which have certain characteristics in common. But this simple interpretation fails to satisfy the theorist when he tries to find definite functional relations, for example, the supply and demand curves of these commodities. The complicated algebraic formulae of modern [economic] theory are evidently built on the assumption that composite commodities have exactly the same definitely measurable dimensions of quality, price, utility, etc., as any of the individual commodities." (Leontief, 1936, p. 39.)

It should hardly be necessary to argue that economics as we know it depends on our ability to measure and theorize about the economic properties of such composite commodities as "real wages" and "producer goods." If one would pay the price of rigor and deal only with homogeneous goods, economic theory and policy based on any level of aggregation would evaporate. Composite commodities are completely integral to economic inquiry in exactly the same way generality is integral to all scientific inquiry. On the basis of similarities we group specific items into categories which allow us to generalize about untested properties and to predict unseen or future events. Whether categories exist in reality or only in the human mind, this assumption is absolutely essential to all rational and organized human thought.

Our problem, then, is that we are attempting to deal quantitatively with an idealized concept rather than more mundanely with a real measurable phenomenon. This is fundamental and endemic to the social sciences generally, although perhaps less serious in economics. Therefore, as a measure of our idealized concept we look for a "proxy" which is a real and measurable phenomenon in some sense congruent with or consequent upon our idealized concept. For example, the concept of "worker militancy" is required for certain theories of wage determination. Since this concept is metaphorical (or even metaphysical) we substitute instead some related series such as "days lost through strikes," the "change in trade union membership" or "number of Communist shop-stewards." None of these series is exactly what we want, and no two would

necessarily agree very closely, but we have managed to measure our idealized concept, however ambiguously.

The indexing situation is rather more optimistic. Following from the purpose and intent of a composite commodity is that it should be a composite of the separate properties of its components, and that these composite properties should behave like the component properties. This was the approach of Fisher (1922), whose work after more than half a century still defines the analogue approach to index numbers.[2] In operational terms, his requirements are that the prices and quantities should (1) be an average of the prices and quantities of the component commodities, and (2) should behave appropriately, that is, pass the several analogue tests devised by Fisher. These requirements are in no sense an artificial constraint, but are rather implied by our original premise.

Having stated what was implicit narrows, but does not eliminate the resulting ambiguity. In the first place, there are a great many types of averages. An average may be either arithmetic, geometric or harmonic, and may be either unweighted or weighted with initial or final weights, or any average of these. Moreover, each simple index may be averaged in any way with any other combination. When several years are to be compared, each result may be a direct comparison or an aggregation of individual adjacent changes. Fisher lists 125 formulae which include virtually every combination of characteristics.

Having defined the range of possibilities, Fisher then eliminates those averages which do not behave appropriately, that is, as would price and quantity series for a homogeneous good. Fisher's famous tests (Allen, 1975, pp. 44–47) define the necessary properties. The first three are basic and somewhat trivial. They include the properties of identity, proportionality and commensurability which are satisfied by almost every formula. The remaining three tests include time reversal, factor reversal and circularity and are more controversial. Time reversal requires that time-reversed indices be reciprocal, i.e. $P_{0t} \times P_{t0} = 1$. No one actually criticizes this property, but it is often ignored. Neither Paasche nor Laspeyres indices have this property, for instance. Factor reversal requires that price and quantity change partition the value change, i.e. $P_{0t} \times Q_{0t} = V_{0t}$, which also defines Fisher's concept of index bias.[3] An indexing formula for which $P_{0t} \times Q_{0t} > V_{0t}$ is said to have an upward bias, and in the opposite case a downward bias. Vartia (1978) mathematically explicates and proves the quantum nature of the empirical biases noticed by Fisher. On the other hand, Samuelson and Swamy (1974) reject Fisher's concept of index bias (p. 567) and the factor reversal test (p. 575), although without explanation.

The easiest way to satisfy time or factor reversal is to "cross" (geometrically average) an index with its time or factor antithesis (i.e. complementary indices satisfying time or factor reversal). The Paasche and Laspeyres indices are both time and factor antitheses, because their geometric average (the Fisher) satisfies both properties, which Fisher called "ideal." It is not unique in this regard, but was preferred by Fisher as being the simplest and having the best economic

[2]See Ruggles (1967) for a brief but comprehensive historical survey (and bibliography) of index numbers.

[3]I.e., total bias $\equiv P_{0t} \times Q_{0t} / V_{0t}$; average bias $\equiv$ (total bias)$^{1/2}$.

interpretation. Every ideal index is unbiased, and every unbiased index is at least close to being ideal.

Circularity requires an ordered partition of index change, i.e. $P_{0s} \times P_{st} = P_{0t}$, and is therefore not defined for inter-country comparisons which have no particular order. It is not satisfied for any economically relevant direct index, but is for every chain-linked index which is of necessity based on ordered data. Fisher particularly disliked chain-linked indices and, therefore, ended up rejecting what he finally called the "so-called circular test," which became a chapter title. The basis of his pique seems to be his use of World War I data (1913–18), which distorted some of his other conclusions as well. Samuelson and Swamy (p. 576) strongly declare for circularity as a basic theoretical necessity. Allen (1975) demonstrates that no statistical expectation exists for an annually chain-linked index to drift from its direct counterpart (pp. 186–188), nor is there any very compelling empirical evidence to suggest that this should be true.[4] Furthermore, every index has to be rebased and linked fairly frequently for purely practical reasons, so there is an obvious temptation to carry this to its logical conclusion and chain on an annual basis. In other words, on an analogue basis one would like to chain annually while in practical terms there is no reason not to.

Our conclusion, then, of the analogue test approach is that, rather than being artificial or arbitrary, it follows naturally from the basic inclusion in economics of composite commodities; therefore, in general, the formula default should be an ideal chain-linked index. This subset of indices does not, however, define a range for some "true" index, which does not exist under the analogue approach. Moreover, this is not to say that other indices are necessarily invalid, but only that they are not logically consistent proxies for composite commodities. And there are special, subjective situations (as we shall see) where another approach might be entirely appropriate.

B. *Considerations from Theory*: "*Exactness*"

The indexing formulae which have managed to elicit some theoretical comment include several which are not as well known as the Paasche, Laspeyres and Fisher. The list of relevant indices comprises a rather small subset of the possible averages, which can be seen in its virtual entirety in Fisher (pp. 467–488):

1.	Laspeyres	weighted arithmetic average using initial expenditure shares
2.	Paasche	weighted harmonic average using final expenditure shares
3.	Jevons	unweighted geometric average
4.	Fisher	geometric average of Paasche and Laspeyres

[4]The Fowler study (1970, 1973 version quoted in Allen, pp. 191–197) is one of the few published empirical tests of this proposition and, based on ten years of British household expenditure data, concluded: "The shorter the time interval over which price indices are computed the closer the Laspeyres and Paasche indices can be expected to be" (p. 15): Unfortunately, our experience with Egyptian foreign trade data was exactly the opposite (1979, pp. 41–50), and it appears that Fowler is guilty of greatly over-generalizing about a result that is probably data-specific (and therefore economic). Allen (p. 188) still would seem to have the last word: "there is no reason to expect that the chain Laspeyres drifts above the direct Laspeyres index nor, equally, that it tends to correct for any propensity for the direct index to run high. Empirical evidence is needed. . . ."

5.	Törnqvist	weighted geometric average using arithmetic average of initial and final expenditure shares
6.	Sato–Vartia	weighted geometric average using the logarithmic average[5] of initial and final expenditure shares

Of the indices listed, only the Fisher and the Sato–Vartia (Sato, 1976) are ideal, although the Törnqvist is unbiased. In all cases the variable being averaged consists of the price or quantity relative (i.e. ratio of final to initial price or quantity) for each commodity, the weights being some function of the value of the commodity in question in the initial and final periods. Every index is, then, a binary comparison of average price or quantity change between two periods (or geographically separate economies). Multiple indices are some combination of binary indices and consist of segments composed of direct indices (each using a common initial period) which are equilibrated at their respective intersections. Therefore, the reference base (= 100) is generally not the same as the weights base (initial period for each computation) which will be different for different segments of a multiple index. The limiting situations are a multiple index which is completely *direct* or one which is rebased and linked every year (i.e. *chain-linked*).[6]

In response to the analogue approach of Fisher, the early modern theorists began to assert themselves:

> "...[A]ll discussions about the "best" index formula, the "most correct" weights, etc., must be vague and indeterminate so long as the meaning of the index is not exactly defined. Such a definition cannot be given on empirical grounds only but requires theoretical considerations." (Frisch, 1936, p. 1.)

They were equally critical of the earliest theoretical approach (associated with Edgeworth) which they described as "merely" stochastic:

> "The assumption is made that any change that takes place in the "price level" ought, so to speak, to manifest itself as a proportional change of all prices. Whatever deviation there is from this strict proportionality must be looked upon as due to other causes than those we think of when we speak of the price level change....
>
> According to this conception, the deviation of the individual price changes from proportionality must be considered more or less as errors of observations. But then the applications of the theory of errors should enable us to determine the underlying proportionality factor....
>
> Thus, the notion of a "price level" here becomes essentially stochastic." (*Ibid.*, p. 3).

Alternatively, "We face the deviations from proportionality and take them merely as expressions for those systematic relations that serve to give an economic meaning to the index number." (*Ibid.*, p. 10). Keynes was the pre-eminent spokes-

[5]The logarithmic average of w_0, w_1 is defined as $(w_1 - w_0)/\ln(w_1/w_0)$.

[6]See Allen, *op. cit.*, for discussion of the whole range of problems associated with multiple indices.

man for this new approach:

> "It follows that [an index] must always be defined with reference to a particular set of individuals in a given situation namely those whose actual consumption furnishes us with our standard, and has no clear meaning unless this reference has been given." (Keynes, 1930/Allen, 1975, p. 7.)

The economic theorists won the argument, although in fairness to the stochastic approach the resulting index is the same if the variable being indexed and the weights are uncorrelated (Bowley, 1911), not an obviously bad assumption.

The outcome of the debate was the development of "exact" indices which are explained mathematically in Diewert (1976) and graphically in Moorstein (1961). The point is to measure the proportional "distance" between two production or utility surfaces (≡ aggregator functions) which gives an exact measure of the change involved subject to the validity of the assumed functional forms. For any homothetic aggregator function there is a unique or "invariant" solution. For the nonhomothetic situation the result varies with the point of comparison, which is the "exact" equivalent of the index number problem. The problem becomes one, then, of matching index formulae with and evaluating aggregator functions, of assessing the plausibility of the homotheticity assumption and the consequences of relaxing it, and, finally, the application of possible restrictions that would produce an invariant result in the non-homothetic case.

Before evaluating the "exact" approach, it is useful to look at the most widely held alternative—the null hypothesis, in other words. This view holds that all indices are "subjective," that change is measured only from a particular arbitrary perspective, and that, therefore, each index is equally valid and equally invalid. Or, to put it another way, validity is not a proper attribute of indices. This would seem to be a bit of an overreaction to the unavoidable imprecision of an index which is, after all, a proxy—an overreaction equivalent to throwing the baby out with the bath water. The only sort of situation where this approach might be useful would be some truly subjective test such as a personal price index where the reaction of the most inflexible participant is of interest, i.e. a Laspeyres index. This would produce a minimum compensatory boundary beyond which everyone could be considered to be better off. This test might have its applications, but in general we are looking for a better measure of a more precise concept and the "exact" approach gives it to us.

There are an unlimited number of uniquely corresponding homothetic aggregator functions and exact index formulae (Samuelson and Swamy, 1974). That is, each homothetic aggregator function implies a unique index formula which is exact, and each index formula implies a unique homothetic aggregator function for which it is exact. For economically relevant homothetic aggregator functions the following correspondences exist:

1.	Jevons	Cobb-Douglas aggregator function (Samuelson and Swamy, 1974)
2.	Fisher	Konüs and Buscheguence (1926) homogeneous quadratic aggregator function, of which the Leontief and linear aggregator functions are a special case (Diewert, 1976)

3.	Törnqvist	Homogeneous transcendental logarithmic (translog) aggregator function and its dual (Diewart, 1976)
4.	Sato–Vartia	Homogeneous constant elasticity of substitution (CES) aggregator function (Sato, 1976), which is a special case of a CES composition of Cobb–Douglas aggregator functions (Lau, 1978)

Given the range of useful aggregator functions, what can we expect of the corresponding index formulae? Assuming homotheticity, any symmetric mean of the Paasche and Laspeyres (e.g. Fisher) will "approximate the true index ... up to the third order in accuracy" (Samuelson and Swamy, p. 582). In general, the logarithmic difference between the Fisher and the Törnqvist is one of the third degree in the deviations of price and value log changes, or very small indeed." (Vartia, p. 292). Finally, the weighting functions of the Törnqvist and the Sato–Vartia (which is their only difference) are approximately equal, the error being "of the second order" in the percentage change of the weights (Sato, p. 224). In general, then, these indices should be fairly good approximations of each other, and in the homothetic case fairly good approximations of the "true" index.

How reasonable is the assumption of homotheticity for an aggregator function? Probably not very good, with the possible exception of production theory which could reasonably specify constant returns to scale (Samuelson and Swamy, p. 577). In general, though, what consequence results from relaxing the homothetic constraint and what alternatives exist? Principally, there no longer exists an exact index, but some range depending on the point of comparison. At worst, then, we are reduced to a proxy for our idealized, no longer measurable concept. If our alternative is a Paasche or Laspeyres index, an exact formula may give a significantly better result which is at least within the range defined by the initial and final true index. From Moorstein (1961), the use of a Paasche or Laspeyres index is equivalent to assuming a plane tangent to and approximating the aggregator surface at the point of final or initial composition. This procedure is not economically plausible and introduces obvious errors of measurement which could be much more serious than the resulting ambiguity of the non-homothetic exact index.

Finally, is it possible in the non-homothetic case to have an exact index number under reasonable economic assumptions? The answer, courtesy of Sato (1976), is yes. The required concept is the Divisia index, which is necessarily true at any instant of time (or for any incremental spatial interval). The instantaneous index is then integrated to achieve the usual interval index which depends both on the aggregator function *and* the time (or spatial) path of change. Since there are an infinite number of paths, there is no longer a unique correspondence between the aggregator function and exact index formulae, but there are infinitely many path-aggregator combinations for which the index is exact. The problem then reduces to that of selecting economically plausible scenarios which will define an index in the general, non-homothetic case.

Initially, Sato derives the Divisia index as a log change index (i.e. geometric) which, in computational terms, implies chain-linking as well for a multiple index (Allen, p. 177). Then he specifies the time path for the indexed variable and its weights, which then implies both the index formula and the associated aggregator

function. If the indexed variable can plausibly be interpolated exponentially between observations, and the weights linearly, the implied formula is the Törnqvist and its associated aggregator function is a non-homothetic translog function. If the index variable and its weights grow exponentially, the implied formula is the Sato–Vartia and the associated aggregator function is the direct and indirect addilog functions of Houthakker (1960) for quantities and prices.[7] While neither associated aggregator function is uniquely associated with its exact indexing formula in the non-homothetic case, each reduces to the uniquely corresponding form when homotheticity is imposed. In conclusion, then, even in the more realistic non-homothetic case it is theoretically possible to specify an exact index formula based on not implausible economic assumptions.

C. *Empirical Results*

To have economic meaning an indexing procedure should be from the subset of formulae passing both analogue and theoretical tests, and which is, therefore, both "ideal" (or at least "unbiased"), and "exact" for some plausible set of economic assumptions. These are not arbitrary constraints, but follow rather strictly from the economics of the indexing situation. The acceptable existing formulae are the Sato–Vartia, the Fisher and to a lesser extent the Törnqvist. Furthermore, the theoretical expectation is that these preferred formulae should approximate one another (and the "true" index if it exists), although generally differing somewhat based on the differing implicit economic assumptions.

To test this expectation, we have computed price and quantity indices for imports, exports and net barter terms of trade from our Egyptian foreign trade data for the period 1885–1961. In all cases the indices are chain-linked for 70 intervals over 76 years (excluding 1940–45). The index results (1885 = 100) are presented in Table 1, and include also for comparison the Paasche, Laspeyres and implicit Törnqvist (factor antithesis of the Törnqvist) formulae. The factor antithetical discrepancies (see p. 27) are presented in Table 2, expressed as a percentage of the average of the Fisher and Sato–Vartia formulae, and our theoretical expectations are not contradicted: the Sato–Vartia and the Fisher are

TABLE 1

EGYPTIAN FOREIGN TRADE INDICES (1961)

(1885 = 100)

	QM	PM	QX	PX	NBTT
PCH	338.27	341.80	147.99	551.02	1.61
TOR	489.91	530.87	190.83	723.23	1.36
SAV	507.46	530.52	192.78	725.78	1.37
FSH	516.17	521.56	193.85	721.77	1.38
(TOR)	507.13	549.52	193.46	733.20	1.33
LAS	787.65	795.87	253.92	945.44	1.19

Note: See Table 2

[7]For a geographical index, of course, the path of change is strictly hypothetical and would involve one economy transforming itself incrementally into another, rather than actually being transformed as happens in one economy over time.

TABLE 2

INDEX DISCREPANCIES (1885–1961)

(Percentages)

	QM	PM	QX	PX	NBTT
SAV-FSH	1.7	1.7	0.6	0.6	0.7
TOR-(TOR)	4.3	4.3	1.4	1.6	2.2
PCH-LAS	87.8	86.3	54.8	54.5	30.5

Note: Q quantity, P price, X export, M import, NBTT net barter terms of trade, PCH Paasche, TOR Törnqvist, SAV Sato–Vartia, FSH Fisher, (TOR) implicit Törnqvist, LAS Laspeyres.

virtually identical, the Törnqvist differs by a few per cent, while the Paasche and Laspeyres are so extremely divergent as to be completely meaningless. The discrepancies can be traced to the underlying formula biases (Table 3), which are the result of very unrealistic economic assumptions in the case of the Paasche and Laspeyres.

TABLE 3

INDEX BIAS[a] (1885–1961)

(Percentages)

	M-total	M-average	X-total	X-average
PCH	−57.0	−34.5	−41.7	−23.7
TOR	−3.4	−1.7	−1.4	−0.7
LAS	132.9	52.6	71.6	31.0

[a]See footnote 2.

Our conclusion, then, may be stated with at least some confidence: *when based on a fairly wide range of plausible quantitative economic assumptions, an index is enormously robust with respect to formula.* Whether we assume any of the usual homothetic aggregator functions, or in the non-homothetic case any time path ranging from linear to exponential, the result is virtually the same. The only theoretical constraint is that we confine ourselves to certain generally plausible economic scenarios. Consequently, based on their rather implausible economic assumptions, it is no longer justifiable to use the Paasche or Laspeyres formulae, except possibly in situations where a subjective rather than an economic result is desired.

II. INDEX CONFIDENCE

The concept of index confidence has existed for almost as long as indices themselves, but both its purpose and statistical technique have varied considerably. In all, three separate problems have been identified in the literature: (1) "instrumental error," (2) errors of measurement, and (3) sampling error, each referring to separate elements of uncertainty in the indexing procedure.

"Instrumental error" was coined by Fisher to refer to the formula "variance," which he actually computed using each of 13 approved formulae to generate an element for his sample (p. 407). This is not a believable statistical measure, but his conclusions were remarkably similar in tone to our own:

> "The "instrumental error"... can be reduced by the right choice of formula so low as to be negligible as compared with the errors from other sources—particularly the assortment of the commodities and their number." (p. 349)

The problem of errors of measurement for indices was introduced by Bowley (1897, 1911), although this important early contribution has led to some confusion. He is misquoted by Mills (1927, p. 241) and even by Allen (p. 246). Bowley's concern is with the effects of errors of measurement only, not with any type of sampling error, and his simplifying assumptions probably assume too much—principally that the weights and the relatives being indexed are uncorrelated (1911, p. 84). If this were true, there would be no difference between a weighted and an unweighted index. His approach, however, is important, although it has yet to be properly developed and applied. He looks at the problem of data confidence (e.g. reporting errors) and proceeds to estimate the uncertainty this introduces into the indexing calculation. Further work remains to be done, but a general result would be extremely useful for a problem which is rather endemic to indexing situations.

By far the largest source of uncertainty, as well as the most discussed aspect of uncertainty, is sampling error. Sampling error is a function of both sample size and the underlying variability of the data. Since there is a cost to sampling, it would of course be desirable to limit the sample to that required by the necessary accuracy of the final result. Conversely, given a certain effort one would like to know how good the result is, in the sense of its being reproducible. That, by the way, is all that we can measure with this approach. We have to assume that the data are accurate and economically meaningful, which are not necessarily good assumptions, especially with aggregate data and the resultant problem of heterogeneity (i.e. quality change). As an added bonus, the underlying variability of the data is itself an important economic indicator: the variances of the price or quantity relatives are a measure of the non-proportionality of change. They are, then, a measure of structural change within the economy, of changing preferences or production possibilities in theoretical terms. These two statistics (as well as the correlation between them, which indicates the nature of the market response) can be used for an extended analysis, but for now they are important in terms of the generality they imply. The less variable the data the more general the resulting index in the sense that it would approximate its (uncomputed) subindices, the resulting error being completely analogous to sampling error.

The standard error for an index may either be computed or simulated. The former would be preferable, being simpler and more general, but the simplifying assumptions required by statistical theory do not always coincide with economic practice. Therefore, we have the option of trying to develop statistical theory to coincide with economic practice, or changing economic practice to coincide with the requirements of statistical theory. As a fall-back, it is always possible to

compute a series of separately sampled indices from the same data and compute the experimental variance. In any case, this latter would be a useful test of any theoretical result.

Before considering alternative statistical assumptions, it is useful to look at the implications of the two methods of data generation. First of all, index data may be homogeneous, experimental data used with predetermined weights derived from survey information. This, for instance, is the method generally used for a consumer price index. Since it is based on individual transactions it gives excellent homogeneity control, and there is no practical limit to the amount of data that may be generated. The small coverage is amplified through the use of stratified sampling, which computes a series of subindices which are then combined to the desired level of aggregation. The problem with this technique is, first of all, the huge effort involved, and secondly, the limitation usually to contemporaneous data. For most indexing problems there do not exist sufficient resources or data to make this a realistic option. When they do, however, the sampling errors derived from experimentally derived data and predetermined weights are straightforward and unambiguous (e.g. Allen, pp. 241–245).

The typical index, on the other hand, is based on published, aggregate data which includes weights. In this case, the total economic activity under consideration is somewhat fictitiously partitioned into descriptive "commodities" (e.g. textile machinery). These aggregate commodities have only average price (i.e. unit value) data, and heterogeneity is an unknowable problem. The advantage is that the data are relatively cheap to use and exist for the past as well as the present. For aggregate data, the sampling problem becomes one of selecting which subset of commodities are to be included in the index. The resulting error is measured against the complete population index which is the standard only in that it has zero variance. Problems of heterogeneity (and poor formula choice) could result in a very poor index in economic terms.

In terms of statistical assumptions there are several options, but there are no realistic sampling options other than simple random sampling. Most statistical theory is in terms of simple random sampling since other procedures are often difficult or impossible to calculate. The element of selection for an index may be either the commodity or the unit of expenditure (e.g. dollar), and each may be sampled either with or without replacement. As a realistic description of economic practice we can exclude sampling with replacement, although Adelman (1958) argues that it would have several advantages. Of the two remaining possibilities neither models reality very well, but we are prepared to argue that they are the best theoretical estimates of the actual indexing variance, and that they provide an upper and lower boundary for the true uncertainty. Sampling by commodity overstates the variance since economists invariably select the largest commodities first, rather than at random, which tend to dominate the final result. Furthermore, the larger commodities are less variable generally, because they represent a larger number of transactions. Conversely, selection by expenditure understates the variance since economists invariably take the entire commodity selected rather than the single dollar of expenditure, so that the k dollars of the designated commodity are actually 1 choice rather than the k choices in the calculation. Sampling by expenditure is not entirely far-fetched, and with replacement is

equivalent to sampling by commodity with a probability of selection proportional to its value.

The mathematics of the two measures of the variance are fairly standard. The population index (for weights, w and variable, x),

$$\theta = \frac{\sum^N w_i x_i}{\sum^N w_i} \tag{1}$$

is our parameter, which is estimated from a sampled subset:

$$\hat{\theta}_c = \frac{\sum^n w_i x_i}{\sum^n w_i}, \tag{2a}$$

or

$$\hat{\theta}_e = \frac{\sum^n x_i}{n}. \tag{2b}$$

If we sample by commodity as in (2a) the index estimate is the ratio of the sum or average of two random numbers, $w \cdot x$ and w, the coefficient of variation of which (Cochran, 1977, p. 154) is:

$$C^2_{\hat{\theta}_c} = \frac{1-f_c}{n_c}(C^2_{wx} + C^2_w - 2r_{w \cdot wx} C_{wx} C_w) \tag{3a}$$

where $f_c = n/N$ and $n_c = n$. Sampling by expenditure as in (2b), yields

$$C^2_{\hat{\theta}_e} = \frac{1-f_e}{n_e} C^2_x, \tag{3b}$$

where C_x is the weighted coefficient of variation for x, the price or quantity relatives, and f_e and n_e refer to covered expenditure rather than number of commodities. To calculate the cumulative interval for a chain-linked index, one can calculate, assuming independence:

$$C^2_{\hat{\theta}_j} = \prod_{j=1}^{t} (C^2_{\hat{\theta}_j} + 1) - 1. \tag{4}$$

Sampling by commodity results in a biased estimate, which Cochran limits and approximates as follows:

$$\frac{E(\hat{\theta}) - \theta}{\sigma\hat{\theta}} \leq C_{\bar{w}} \qquad \text{(p. 162)}$$

$$\leq C_{\bar{w}} \frac{C_{\bar{w}} - r_{wx \cdot w} C_{\bar{w}\bar{x}}}{C_{\hat{\theta}}} \qquad \text{(p. 161).}$$

The values for the limit tend to be high for our Egyptian import and export indices, occasionally exceeding 50 percent, but the approximation averages only about 5 percent. We can conclude, then, that bias is not a problem in our indices, although it would be at some lower coverage.

To calculate the limits of our index variance, we divide our commodities into two strata, those with price and quantity relatives and those with value data

only, and assume that both strata have the same variance. Using equation (3b) and sampling the entire covered stratum, the resulting index variance is extremely small. Since $C_x < 0.5$, $f_e > 0.7$ and $n_e > 4 \times 10^6$ (Hansen and Lucas, 1979, Tables A.1.a.1, A.1.c.1, A.2.a.1, A.2.c.1), $C_{\hat{\theta}e} < 0.00014$ which by equation (4) would accumulate to less than 0.0013 for 1885–1961. In other words, the lower limit is so low as to be negligible. The upper limit can be calculated from equation (3a), the components of which are presented in Tables A.1.b.2 and A.2.b.2, and the final calculations (95 percent confidence) are presented in Tables A.1.b.3 and A.2.b.3 and summarized in Table 4:

TABLE 4
CUMULATIVE ERROR RATIO
(95% confidence)

	QM	PM	QX	PX
1885–1961	0.62	0.33	1.07	0.39
(annual average)	(0.069)	(0.038)	(0.115)	(0.044)

These estimates of sampling error appear quite high, in contrast to the lower bound which was extremely low. The current statistical situation, then, does not seem to be particularly helpful in terms of measuring index confidence.

Historically, there is not much to compare. Bowley and Mills calculated confidence intervals for unweighted indices, which is fairly simple. Mills (p. 241) calculated a confidence interval for various weighted indices, but he used a doubly erroneous formula from Bowley (1911). Lipsey (1963), in his classic study of American foreign trade, also computed several standard errors, but seems to have mixed his statistical assumptions. His principal formula (p. 379), in our notation would be:

$$C^2_{\hat{\theta}} = \frac{1 - f_e}{n_c} C^2_x,$$

which substitutes the commodity (n_c) for the expenditure (n_e) sample size. The resulting standard errors are numerically believable, but it is not clear from his discussion what the statistical basis is for computing it this way.

The simulation results that exist tend to minimize the importance of sample size. Both Mitchell (1915, pp. 44–71) and Fisher (1922, pp. 336–340) ran simulations for various sample sizes (including one of 3 for Fisher!) and concluded that index confidence improved only very slowly, compared to theory, when sample size was increased. Their methodology is completely obscure, and their result, based on very few simulations (5 for each), must be considered intriguing rather than definitive. Since this experiment has not been conducted for over 50 years, and never under very rigorous conditions, this would seem a useful approach given what appears to be a theoretical impasse.

BIBLIOGRAPHY

Adelman, I., A New Approach to the Construction of Index Numbers, *Review Econ. & Stat.*, 40 (3), August 1958.

Allen, R. G. D., *Index Numbers in Theory and Practice*, Aldine, Chicago, 1975.

Bowley, A. L., Relations Between the Accuracy of an Average and That of Its Constituent Parts, *J. Royal Stat. Soc.*, 60 (4), December 1897.

———, The Measurement of the Accuracy of an Average, *J. Royal Stat. Soc.* 75 (1), December 1911.

Cochran, W. G., *Sampling Techniques*, John Wiley & Sons, New York, 1977.

Diewert, W. E., Exact and Superlative Index Numbers, *J. Econometrics*, 4 (2), May 1976.

Fisher, I., *The Making of Index Numbers*, Riverside Press, Cambridge, 1922.

Fowler, R. F., *Some Problems of Index Number Construction*, Studies in Official Statistics, Research Series, No. 3. HMSO, London, 1970.

———, Further Problems of Index Number Construction. Studies in Official Statistics. Research Series, No. 5. HMSO, London 1973.

Frisch, R., The Problem of Index Numbers, *Econometrica*, 4 (1), January 1936.

Hansen, B. and Lucas, E. F., Egyptian Foreign Trade, 1885–1961: A New Set of Trade Indices, *Journal Eur. Econ. Hist.*, 7, 1978.

———, A Statistical Approach to Index Numbers, unpublished, 1979.

Houthakker, H. S., Additive Preferences, *Econometrica*, 28(2), April 1960.

Lau, L. J., On Exact Index Numbers, *Review Econ. & Stat.*, 61 (1), February 1979.

Leontief, W., Composite Commodities and the Problem of Index Numbers, *Econometrica*, 4 (1), *January* 1936.

Lipsey, R. E., *Price and Quantity Trends in the Foreign Trade of the United States* (N.B.E.R.), Princeton University Press, Princeton, 1963.

Mills, F. C., *The Behavior of Prices*, N.B.E.R., New York, 1927.

Mitchell, W. C., *The Making and Using of Index Numbers*, U.S. Government Printing Office, Washington, 1915.

Moorstein, R. H., On Measuring Productive Potential and Relative Efficiency, *Quart. J. Econ.*, 75 (3), August 1961.

Ruggles, R., Price Indexes and International Price Comparisons, *Ten Economic Studies in the Tradition of Irving Fisher*, John Wiley & Sons, New York, 1967.

Samuelson, P. A. and Swamy, S., Invariant Economic Index Numbers and Canonical Duality: Survey and Synthesis, *American Economic Rev.*, 64 (4), September 1974.

Sato, K., The Ideal Log-Change Index Number, *Review Econ. & Stat.*, 58 (2), May 1976.

Vartia, Y. O., Fisher's Five-Tined Fork and Other Quantum Theories of Index Numbers, in W. Eichhorn, *et al.*, *Theory and Applications of Economic Indices*, Physica-Verlag, Würzburg, 1978.