Florida Journal of Educational Research Fall, 1987, Vol. 29, No. 1, Pp. 5-28.

A Comfort Index for Statistical Inference-Making

James K. Brewer
Florida State University

ABSTRACT. An index that measures the degree of comfort a researcher has relative to a statistical inference application is proposed. The index is calculated by selecting and weighting components inherent in the conduct of hypothesis tests and confidence intervals, and its values range from 0 to 1. Illustrations of calculating and interpreting the index using a published research report and the comparability of indices produced by several researchers from one research report are described. Potential applications and limitations of the index are discussed.

Those of us who function as bridges between the theoretical world of statistics and the statistical practitioner are aware of the many factors to be considered in choosing, applying and defending statistical inferential techniques. In our teaching, consulting and research we repeatedly list and discuss these factors to provide some level of "comfort" for the user in applying statistical inference techniques to practical situations. Rules-of-thumb and robustness arguments are commonly invoked to ease the strain associated with questionable assumptions, inadequate data and bothersome interrelationships.

The proposed "comfort index" for statistical inference-making incorporates many of the crucial ingredients, conditions, assumptions and

interpretations of statistical inference. The purpose for the index is to provide the researcher and readers of research with an overall, general and numerical indicator of "how good they feel" about the inference made. A by-product of the index is its components which serve as a guide to the proper application of statistical inference procedures. Inference-making is restricted here to the use of hypothesis testing (HT) and parameter estimation (confidence intervals, denoted CI) techniques.

This index-producing activity, like most statistical inference, is subjective because the researcher makes judgments about what components to include and the relative importance of each. Since researchers do not agree what components should be considered or on their relative merits, researcher-dependent weightings of the inference-making components are used.

The index proposed has several possible applications for behavioral researchers. One possible use is in reviewing, critiquing and evaluating statistical inferences in papers submitted for publication by research journals. The index would provide a numerical summary of how the paper reviewer assessed the statistical inference made. Likewise, the statistical inference portions of previously published articles could be evaluated with the index, particularly when comparing several articles all of which use the same inference procedures (e.g., meta-analyses preparations). Researchers might also use the index to decide which alternative statistical procedures provide them with the most comfort relative to assumptions, sampling restrictions, and any number of other situation-dependent aspects of statistical inference. Another application could be appraising the inferences made through statistics in theses and dissertations in graduate training programs. This latter use, however, might entail consideration of different components and weights than those used in published works since dissertations include more detail than articles submitted to journals.

These applications as well as the choice of appropriate components on which the index is structured, are functions of the audience who is to interpret the index and of the intent of the index user. Some journals presume more than a modest level of statistical sophistication for readers, whereas others assume their readers have little or no statistical expertise. Likewise, the amount of information, justification and detail expected in statistical reports varies in the behavioral science literature, at times independent of the expected sophistication of the reader.

In order to help offset this variability in expertise and detail, the proposed index enables index users to choose an appropriate set of components, weight them according to their expectations, and produce a value between 0 and 1 to reflect their degree of comfort with the statistical inference being evaluated.

If one were to agree that the above applications provide opportunities for evaluation of statistical inferences, why would an index be more useful than expressions such as "appropriate", "adequate", "satisfactory", "weak", or "acceptable"? Besides the apparent natural penchant for humans to produce indices for everything, this index causes the users to consider and display the components of their evaluation and the weights they chose for each in producing the index. These declarations should clarify the bases for judgments made on any statistical inference application.

Subsequent sections of this paper describe the components of the index, the weighting scheme to be used, and an application using a recent article published in the behavioral sciences literature.

The Components of a Comfort Index

To develop a reasonably comprehensive set of components to be considered for inclusion in the comfort index, a rational beginning point must be defined. Although it is known that internal and external validity issues (Campbell and Stanley, 1963) as well as other measurement concerns are vital to the interpretation and generalizability of any statistical inference, it is assumed these issues have been resolved satisfactorily by the researcher. Further, it is presumed the statistical design of the study is proper and that either HT or CI (or both) is appropriate for the design.

To organize the components of the index, inference components are divided into two possibly overlapping categories: a priori (before collecting sample data) and post hoc (after collecting sample data). Examples of these two types of components are discussed in the following paragraphs.

A Priori Components

In any HT or CI application there are several components of each technique to consider prior to collecting data. These include statistical model and assumptions, sampling issues, piloting, and practical concerns such as cost and nature of the audience benefiting from the analyses.

One of the first questions asked by a researcher after establishing a research design is, "How much data do I need?" There is no universal answer to this question since the amount of data depends on the inference technique chosen, assumptions which can be made, and cost considerations. Sample size is situation-and assumption-dependent and always results from subjective judgments made by the researcher in the context of the analysis conducted.

Adequate Sample Sizes for Hypothesis Testing. Regardless of the hypothesis testing technique

ultimately selected or the assumptions made, there are some essential determiners of sample size that cannot be ignored. Detailed discussion of these essentials are provided by Brewer (1986), Cohen (1977), Kirk (1978) as well as other sources, but briefly they are:

1. Effect size (ES): The researcher's opinion of what a minimally important difference is given that H_0 is false.

2. Alpha: The researcher's judgment of the probability of making a Type I error, i.e., rejecting H_0 when H_0 is true.

3. Beta: The researcher's judgment of the probability of making a Type II error, i.e., not rejecting H_0 when H_0 is false to the degree specified by ES. Power, which is 1-beta, is thus the probability of rejecting H_0 when H_0 is false to the degree specified by ES. Clearly, alpha has definition only when ES = 0 and power has definition only when ES$\neq$ 0.

In most situations the population variance is also a determiner of sample size, but often ES is expressed as a function of the standard deviation (Cohen, 1977), thereby incorporating it into ES and eliminating variance from consideration. If it is required and known, then is should be included, but it will not be discussed further here.

Formulas (Brewer, 1986) and tables (Cohen, 1977) for approximate minimum sample sizes for a variety of parametric and nonparametric tests are available and all use in one form or another, ES, alpha, and beta.

Adequate Sample Sizes for Confidence Intervals. Confidence intervals estimate parameters with intervals and as such do not necessarily involve the rejection or nonrejection of hypotheses. Thus, there is no ES, alpha, or beta even though the underlying, data-related assumptions for CI are identical to those in hypothesis testing. The determiners of an adequate

sample size for confidence intervals are:

1. Precision (d): The maximum amount of difference between the parameter and its estimate which is important to the researcher. The word *accuracy* is often used interchangeably with precision, but accuracy is generally reserved for unbiased estimate and parameter differences (Cochran, 1963).

2. Confidence: The researcher's opinion of the probability of "capturing" the parameter of interest with intervals having the specified precision, i.e., when the estimate and the parameter are within d of each other. The confidence is usually expressed as a percentage of the form (1 - e) 100% where e is the probability of intervals failing to capture the parameter, given d.

Confidence intervals, like hypothesis tests, may be used in both parametric and nonparametric situations even though the latter use seems like a contradiction of terms. What is generally meant by *nonparametric* in confidence intervals parlance is that the underlying distribution is not known, i.e., nonparametric here means distribution-free. As in hypothesis testing, variance is often incorporated into the expressions for d and will be so treated except when variance is a function of the parameter itself. Formulas (Brewer, 1986; Cochran, 1963) and tables to approximate minimum sample size for confidence intervals for several different parameters are available and all require consideration of precision and confidence.

Researchers who wish to apply both HT and CI techniques in a single situation must, therefore, consider these five elements to arrive at a minimal sample size satisfactory for both techniques. Some, however, shortcut this process by equating values like alpha and e which is done primarily so that confidence intervals *may* be used to test hypotheses *given* the sample(s). Since the object of this section is to describe the determiners of adequate sample size for

each of the techniques, the five values will be kept conceptually distinct, even though numerically some may be set equal.

Assumptions for Hypothesis Testing and Confidence Intervals. With the obvious exception of differences between HT and CI concerning the purposes of each technique, there are no assumption differences between the two techniques. For example, determining an adequate sample and testing hypotheses concerning μ requires the same statistical assumptions as those to determine adequate samples and to calculate confidence intervals on μ. The same applies to any parameter or combination of parameters, even though formulas for estimating sample size for each may be considerably different and produce different sample sizes.

The type and number of assumptions for inference-making vary and constitute the major distinction between parametric and nonparametric methods. Parametric tests such as the t-test on means require the most stringent set of assumptions whereas nonparametric counterparts such as the Wilcoxon Rank Sum test require fewer and less restrictive assumptions. This observation about assumptions also applies to their confidence intervals. All statistical texts discuss the assumptions inherent in a variety of HT and CI techniques. The following categories of assumptions will prove helpful for developing a comfort index in subsequent sections.

1. Inescapables: Except in extremely rare circumstances, these are the assumptions present in any HT and CI situation, whether parametric or nonparametric, irrespective of other assumptions made. They are the assumptions of (a) randomness of the sample and (b) intrasample (within sample) independence. Ironically, there are no known general statistical tests for these two assumptions. Comfort relative to both these assumptions is gained through the design of the study as well as the sampling process, and satisfying one does not guarantee the other.

2. Situation dependent: These assumptions vary from one inference situation to another and may be classified into two not-so-mutually-exclusive categories.

a. Testables: Those assumptions for which some statistical test is available as a check on the reasonableness of the assumption are termed "testables." Two of the most common are the homogeneity-of-variance and the normality assumptions with the F-ratio (or F max) used for the former and the chi-square or Kolmogorov-Smirnov tests for the latter. Ironically, it may take larger samples to test some assumptions adequately than to conduct the desired hypothesis test due to small effect size and high power concerns.

b. Untestables: The assumptions of known variance, interval (ordinal, nominal, ratio) scale, continuity of variable, and intersample independence are examples of assumptions (other than the inescapables) for which no statistical test is available. The exception to this is the intersample independence assumption which could involve a test of association, but it is included as an untestable because this is generally how the assumption is treated. Some of these assumptions, such as interval scale and continuity of the underlying variable, may be mostly conjecture on the part of the researcher while others (intersample independence) can be supported through proper research designs.

Post Hoc Components

After the sample is selected, the researcher applies the chosen inferential technique to the data. In so doing there are several issues the researcher must consider and some calculations that should be made. It is here that some assumptions may be tested and/or data scrutinized for reasonableness of assumptions, if this was not previously done through pilot data in the a priori stage.

The first calculation in hypothesis testing is the test statistic which is a summarization of the data given H_0 is true. The p value, if it can be directly calculated, is a function of the test statistic. In addition, the table (critical) values can be determined (with degrees-of-freedom where appropriate).

If p is available, a direct comparison should be made between it and alpha to reach a decision. If p is not available, then the comparison should be made between the test statistic and the critical value to reach the same decision.

In confidence intervals, the interval would be calculated keeping in mind the precision and confidence discussed in a priori components. In the event that H_0 is to be tested using the confidence intervals, the decision to reject H_0 (or not) would be made.

Once a statistical test (or series of tests) is conducted or confidence intervals are calculated, there are issues and procedures to be considered by the researcher. There are post hoc indicators of effect that could be entertained and applied to the data for most hypothesis tests. The well-known omega-squared, eta-squared and other proportion-of-variance estimates are some possibilities. Such indicators provide information on the effect size from a post data collection perspective and are the basis of most meta-analysis studies (Hedges and Olkin, 1985), even though some translation from proportion of variance to other commonly used indicators might be necessary.

Follow-up, or multiple range tests, are included in this section since they are hypothesis tests following some main or overall test. For confidence intervals, comparison of actual interval widths with a priori precision would be considered post hoc and a practical importance issue.

The researcher's interpretation of hypothesis tests or confidence intervals is a major post-data activity and is probably the most important from the viewpoint of readers of research. No report would be complete without the researcher's statement about the meaning of the statistical procedure outcome and its practical importance. It is here that nonsense expressions such as "highly significant", "approaching significance" or "the parameter falls within the calculated interval with .95 probability", often creep in and serve to confuse and mislead the reader. Such myth and misconception examples are given by Brewer (1985); however, these expressions should be avoided in research reports or discussions because they add nothing at best, and at worst they imply an untruth. Researchers could state in post hoc discussions what they think the statistical results mean, keeping in mind the fairly restrictive statistical constraints of the procedures used. To go beyond such statements in statistical interpretation is a questionable form of limb-walking.

A Glossary of Statistical Inference Components.

These previous examples of a priori and post hoc components are not intended to be viewed as exhaustive. A glossary of components is provided in Table 1 to extend the list of statistical inference components and include some nonstatistical concerns which indirectly affect a researcher's view of the importance of some statistical components.

Even if this glossary were complete from every standpoint, there would still be combinations of components and conditions that would materially alter the importance of the listed components. For example, a researcher may believe that homogeneity-of-variance is a modestly important assumption, but it might become more important when coupled with knowledge of non-normality and grossly unequal and small samples. Likewise, the invocation of a rule-of-thumb by a researcher for some assumption may offer modest comfort to the reader unless the appropriateness of

Table 1 Alphabetical Glossary of Statistical Inference Components

A Priori (Before Sample Collection)	Post Hoc (After Sample Collection)
alpha	critical (table) values
assumptions	decisions made
central limit theorem	estimate calculations
distributional approximations	inspection of data
homogeneity-of-variance (or regression)	coding errors
intersample independence	multicollinearity
intrasample independence	outliers
model specification	interpretation of results
nature of variable (continuous or discrete)	measures of effect
normality (univariate and multivariate)	p values
randomness	scatter/residual plots
robustness	test statistics
rules-of-thumb	tests of assumptions
scale of measurement	
symmetry	
confidence level	
confirmatory/exploratory	
effect size	
error rate inflation	
hypotheses to be tested	
parameters estimated	
pilot study	
power	
precision (accuracy)	
relative efficiency	
sample size	
sample type	
Nonstatistical Concerns	
costs	
custom/traditional/folklore	
experience and expertise of researcher/user	
practical value	

such a rule is not justifiable.

Rather than attempt to list all possible combinations of components in Table 1, possible scenarios for a few of the components are included in Table 2 as if they were reported in a published manuscript. These non-exhaustive examples indicate ways the components could be viewed to provide differential levels of comfort to the user. Although the component examples are alphabetically listed in Table 1, the alternative scenarios in Table 2 are presented in no particular order since the reporting situation dictates their relative importance. The "worst case" scenario for each component would be its complete absence from a report when it should be considered. There is not necessarily a "best case" scenario; there are only alternatives that provide lesser or greater information and comfort. What constitutes an "appropriate" component will be left to the discretion of the reader/researcher/user who is considering the statistical technique. In preparing a report, the audience, customs, tradition and purposes of the report are a few of the determiners of appropriate components. In other situations, (e.g., comparing several statistical techniques) some components might be omitted since they are held in common by the techniques. The examples in Table 2 demonstrate that the rather formidable list in Table 1 will be reduced to a smaller subset of components for a given situation.

Incorporating and Weighting the Components of a Comfort Index

The weights assigned to separate components are the key to producing any index, although such weights may be determined by the user as in an extensive questionnaire (e.g., indices such as the Cornell Medical Index (Brodman, Erdman, Lorge, & Wolff, 1949)). Indices of association, (e.g., IQ, anxiety, attitude, aptitude, performance) all, to greater or lesser degree, weight or restrict components or variables usually with some fabricated single score or

Table 2 Possible Reporting Scenarios for Some Statistical Inference Components

Components	Scenarios
alpha	stated; used to determine sample size; presented with p value; set equal to 1-confidence
confidence level	stated; used to determine sample size; set equal to 1-alpha.
decisions made	reported; implied by presence of p and alpha.
effect size	stated; used to determine sample size; compared with post hoc effects.
homogeneity of variance (regression)	stated; tested with appropriate test; robustness argument given; residual plots displayed.
hypothesis	both H_0, H_a explicitly stated; one or both implied from design.
measures of effect	omega-squared, eta-squared or some proportion-of-variance measure reported; sample differences displayed; practical value of results given.
normality	mentioned; tested with goodness-of-fit; central limit theorem invoked; implied by nature of instrument and design; robustness argument given.
p value	reported with alpha; reported with test statistic(s).
power	stated; used to determine sample size; discussed with decisions made.
precision (accuracy)	stated; used to determine sample size; compared with interval widths.

Table 2 (Continued)

Components	Scenarios
randomness	stated; details given on how accomplished.
sample size	approximated from consideration of alpha, power, effect size, precision and confidence; stated without justification; rule-of-thumb invoked.
test statistic(s)	reported; implied by test; reported with p value.
tests of assumptions	utilized; utilized with literature justification.

measure being the result. The proposed comfort index is no exception. Like other indices or scales, it serves to define what it purportedly measures, in this case *comfort* associated with statistical inference.

The relative rather than the actual size of weights is crucial in developing an index. Therefore, weights reflect the relative importance of components/conditions as they contribute to statistical comfort. For example, a study in which great care is taken to assure a random sample while no attention is given to assuring intersample independence when required has, in effect, given a heavier weight to "randomness" than to "intersample independence." Similarly, a higher weight (rating) is given to homogeneity-of-variance when equal sample sizes are used, the distribution is normal, and a test of homogeneity-of-variance failed to reject with a modest alpha level than when the researcher merely assumed it or never mentioned homogeneity-of-variance.

To compute the comfort index, each statistical component (hereafter referred to as an "item") listed

in Table 1 that is judged as appropriate for consideration will be assigned ratings from 0 to 5. Zero means either the user viewed the item as unimportant or had minimal comfort with the way the item was used/discussed. The zero rating could also reflect that the item, though appropriate for consideration, was not entertained. The rating of 5 means either the user viewed the item as being of maximum importance or had maximum comfort with the way the item was used/discussed. The intervening ratings 1,2,3,4 reflect either the amount of importance of the item or the levels of comfort with the way the item was used/discussed.

Ratings across some items may be dependent. This is unavoidable, since heavily weighting one item may minimize the importance of another. Ratings in these cases must be conditional ratings based on the presence or absence of information relative to the statistical application.

The Comfort Index Rating Form (CIRF)

The CIRF (Table 3) may be used in an evaluation or appraisal of a completed statistical analysis or as a guideline or checklist in deciding on a particular technique or set of techniques. The user simply scores each item appropriate for the technique considered using the ratings 0,1,2,3,4, or 5. Recall that an *appropriate* item is one the user thinks should be considered for the technique under scrutiny. An appropriate item should be rated zero when it should have been reported but was not; Inappropriate items are left blank. (Table 3 includes ratings for an appropriate set of items from the illustrative example to follow).

After ratings are assigned, the user sums the maximum possible ratings (5) of the appropriate items on the CIRF. Denote this value as T. The actual ratings given to those items are also summed and denoted as A. The comfort index, C, is defined as $C = A/T$. In essence, the comfort index is the ratio of

the sum of actual ratings for appropriate items and the maximum possible sum of ratings on those same items.

An Illustration of the Comfort Index

To see how a reader of research would use the comfort index, consider the following extraction from a recently published paper in the behavioral science research literature. If a researcher were using the index to select a statistical procedure, the process would be very similar to the one described for readers except the components would by hypothetical rather than actual. In this example, the index indicates the reader's comfort in the reported statistical application.

> "The sample consisted of 75 (38 males and 37 females) second grade, middle class students from a midwestern school district. All students were randomly assigned to three conditions stratified for sex and ability level. The dependent variable was student achievement. The ability level of students was determined by...scores on the Gates-McGinnis Standardized Reading Test. The top 1/3 of the students were classified as high ability, the middle 1/3 as medium ability, and the bottom 1/3 as low ability. A 3 X 3 ANOVA was run with the three ability levels and the three experimental conditions as the level of analysis. The overall F tests indicate that there was a significant condition effect on the post achievement test, $F(2,68) = 295.11$, $p < .01$. Newman-Keuls post hoc comparisons revealed that students in the structured-oral discussion cooperative conditions scored higher than did students in the other two conditions, and students in the unstructured-oral-discussion cooperative condition scored higher than did the students in the individualistic condition. This was true for high; medium and low-ability students. These results clearly indicate that...achievement can be increased by structuring the oral interaction of

Table 3 Comfort Index Rating Form (CIRF)

blank the item* is not appropriate for the inference considered

0 the item is appropriate but as used or reported provides minimal information and comfort

5 the item is appropriate and as used or reported provides maximum information and comfort

Rating	Item	Rating	Item
1	alpha	___	nature of variables, factors
_	central limit theorem	0	normality
_	confidence level	2	p value
_	confirmatory/exploratory intent	___	parameters estimated
5	critical values	___	pilot study
5	statistical decisions made	0	power, i.e., 1-beta
_	distributional approx.	___	precision (accuracy)
0	a priori effect size	4	random selection, assignment
_	error rate inflation	___	relative/power efficiency
_	confidence interval values	___	statistical robustness
0	homogeneity-of-variance (regression)	___	statistical rules-of-thumb
2	statistical hypotheses	2	adequacy of sample size
0	inspection of data (outliers, etc.)	5	sample type (stratified, etc.)
5	interpretation of statistical results	___	scale of measurement used
3	intersample independence	___	scatter/residual plots
0	intrasample independence	___	symmetry of distributions
0	post hoc measures of effect	5	inferential test statistic(s)
___	model specification	0	tests of assumptions

* an item is a component for possible consideration in inference-making

> students learning collaboratively. There are practical as well as theoretical implications of the results of this study. The careful structuring of...may considerably increase the efficacy of cooperative learning procedures."

The journal in which this report appeared, the purpose of the study and the intended audience permits the selection of an appropriate set of items. This writer's opinion of an appropriate set of items, his perceived reporting scenarios, and his ratings assigned are included in Table 4. Another reader of the same article could glean a different subset of appropriate items as well as view them as reported under slightly different scenarios, but this is of no consequence in illustrating the use of the index.

Given the 19 components/scenarios in Table 4, ratings were assigned using the ranges shown on the CIRF. Knowing that the maximum rating possible on each of the 19 items is 5, the total possible score, or T, is 95. The actual score on the 19 items (A) is 39 providing a comfort index score (C = A/T) of .41. (These same ratings appear on the CIRF in Table 3.)

The reader could assign other ratings to these items of concern or choose and weight other items producing their own comfort index for the statistical technique used. There is no reason to expect different readers to produce identical ratings for any single item or group of items, but agreement among researchers on such matters is worthy of investigation.

Compatibility of Researchers' Ratings

A preliminary and limited investigation of researchers' agreement on an appropriate set of items and the value assigned to each item was conducted by the author. Two groups of individuals participated in the study. One group, denoted ES, was composed of respondents to a mailout request based on a random sample of 52 members of the AERA Educational

Table 4 Appropriate Items from the CIRF, Scenarios, Ratings for Items, and the Comfort Index

Rating	Appropriate Items	Scenario
1	alpha	not reported, implied
5	critical value	not reported, implied
5	decisions made	reported
0	effect size	not reported
0	homogeneity-of-variance	not reported
2	hypothesis to be tested	not reported, implied
0	inspection of data	not reported
5	interpretation of results	reported
3	intersample independence	not reported, implied
0	intrasample independence	not reported
0	measures of effect	not reported
0	normality	not reported
2	p value	implied, not specified
0	power	not reported
4	random assignment	reported
2	sample size	reported
5	sample type	reported
5	test statistic	reported with degrees of freedom
0	tests of assumptions	not reported

Appropriate items selected from CIRF = 19
T = (5) 19 = 95
A = 39
Comfort Index (C = A/T) = .41

Statisticians Special Interest Group, 1986. The other group, denoted SS, included graduate students completing two courses (EDF 5401 and EDF 5402) at Florida State University in the Fall semester, 1986, who volunteered to rate the items of the CIRF. EDF 5401 is a course in the general linear model and EDF 5402 is a course in the analysis-of-variance with neither course being a prerequisite for the other. No students were in both courses. Responses from both groups were anonymous.

The members of each group read the same statistical abstract from a published article and rated the 36 items using the CIRF. The ES group was told the article had been published in The American Educational Research Journal, and the students were told the article had been published in the behavioral science literature.

Summary data on these two nonrandom samples are shown in Table 5. The average C values for the ES and SS groups were respectively .44 and .41 with standard deviations of .15 and .23. The range of C values for the ES group was from .19 to .78, and for the SS group it was from .04 to .73. A Spearman's rank correlation coefficient between the two groups was calculated on the average ratings per appropriate item. For this analysis an item was included if it was judged appropriate by at least one individual in each group; since all 36 items met this criteria all items were used in the correlational calculation. The Spearman's correlation was $r = .82$.

Discussion

Some of the limitations of such a comfort index are quite apparent. Chief among these is the subjective judgment required to establish an appropriate set of items and to assign ratings to each. Another limitation is the difficulty of teasing out the set of items from research reports and publications that contain varying amounts of information, assumptions

and implications. In many practical situations, multiple variables are entertained, several procedures are used, and alternate statistical techniques applied. Seldom is there an overt explication or discussion of all relevant components, but rather bits and pieces of information are scattered throughout the report with implications or unstated assumptions made relative to statistical inference issues. Additionally, some reasonable and appropriate techniques such as sequential analysis, causal modeling, factor analysis, complicated ANOVA, and nonparametric methods might be difficult, if not impossible, to dissect sufficiently to apply the comfort index. Finally, restricting the index to hypothesis testing and confidence intervals deletes from consideration such topics as point estimation and Bayesian inference.

As if these limitations were not enough, questions like those that follow remain to be answered.

1. What constitutes a sufficiently large index value?

2. Would different researchers produce similar index values given identical situations?

3. Does the index produce reliable values?

4. How should the index value be interpreted and applied?

5. Would differential upper limits on the ranges of ratings be feasible and useful?

Table 5 Frequency and Relative Frequency of Ratings of CIRF Items*

Ratings	Groups ES (n = 14)	SS (n = 22)
0	96 (.18)	215 (.27)
1	46 (.09)	49 (.06)
2	32 (.06)	45 (.06)
3	51 (.10)	52 (.07)
4	60 (.11)	45 (.06)
5	38 (.07)	101 (.13)
Blanks	217 (.40)	285 (.36)

*Note that the total ratings (and blanks) per group over all items is (7)(36)

As is the case with most good questions, the answers to these are not apparent and/or depend on yet more subjective judgments by the reader. Some of the questions might even be researchable. What does seem to be apparent is that the index includes many components of concern in hypothesis testing and confidence intervals, provides relative weights for these components, and produces values between 0 and 1 inclusive such that larger values imply greater comfort than smaller values.

Given that the index indicates the degree of comfort in using statistical analyses, there are several features which encourage its use. One feature is that researcher agreement on an appropriate set of statistical components (or their weights) is not necessary to apply the index. Another feature is that

several indices could be calculated, each on a different subset of components (e.g., separate indices for a priori and post hoc components). As long as the user and interpreter are both aware of the subset of components considered, the possibilities are quite extensive. A third, more complex feature is that users of the index could develop their own set of upper limits for the ratings (weights) dependent on special circumstances and discipline emphases. Finally, negative weights could be introduced to reflect a dampening of comfort in the presence of some components. For example, a modification of the index could be made to provide for the negative impact on the reader of misconceptions or misleading statements present in a statistical report.

References

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? Journal of Educational Statistics, 10(3), 252-268.

Brewer, J. K. (1986). Introductory statistics for researchers. Minneapolis, MN: Burgess.

Brodman, K., Erdman, A. J., Lorge, I. and Wolff, H. G. (1949). The Cornell medical index. An adjunct to medical interview, Journal of American Medical Association, 140, 530-534.

Campbell, D. T., and Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Stokie, IL: Rand McNally.

Cochran, W. G. (1963). Sampling Techniques, (2nd ed.) New York: Wiley.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. (New ed.). New York: Academic Press.

Hedges, L. and Olkin, I. (1985). Statistical methods for meta-analysis. San Francisco: Academic Press.

Kirk, R. E. (1978). Introductory statistics. Monterey, CA: Brooks/Cole.

Brewer, James K., Professor, Educational Research, 307 Stone Bldg. Florida State University, Tallahassee, Florida 32306