Survey Methodology, December 1995
Vol. 21, No. 2, pp. 117–124
Statistics Canada

Design Effects for Correlated $(P_i - P_j)$

LESLIE KISH, MARTIN R. FRANKEL, VIJAY VERMA and NIKO KAĆIROTI[1]

ABSTRACT

We present empirical evidence from 14 surveys in six countries concerning the existence and magnitude of design effects (defts) for five designs of two major types. The first type concerns deft$(p_i - p_j)$, the difference of two proportions from a polytomous variable of three or more categories. The second type uses Chi-square tests for differences from two samples. We find that for all variables in all designs deft$(p_i - p_j) \cong [\text{deft}(p_i) + \text{deft}(p_j)]/2$ are good approximations. These are *empirical* results, and exceptions disprove the existence of mere analytical inequalities. These results hold despite great variations of defts between variables and also between categories of the same variables. They also show the need for sample survey treatment of survey data even for analytical statistics. Furthermore they permit useful approximations of deft$(p_i - p_j)$ from more accessible deft(p_i) values.

KEY WORDS: Design effects; Survey sampling; Sampling errors.

1. DESIGN EFFECTS FOR ANALYTICAL STATISTICS

We explore the existence and the magnitudes of design effects for some special analytical statistics based on data from survey samples. The investigation is both methodological and empirical, with data from several different surveys with different variables and from contrasting populations, hence subject to the risks of inconsistent empirical results. We often hear and read that probability sampling, while necessary for descriptive surveys, is not necessary for analytical surveys. In "Four Obstacles to Representation in Analytic Studies" one of us wrote that "In addition to those four real obstacles, we also encounter another, which is more artificial, in the denials of the need for representation" (Kish 1987, Section 2.7). Sampling investigations show that complex probability selections, especially clustered sampling, have no appreciable influence on descriptive statistics (like means and regression coefficients), but can have drastic effects on inferential statistics, like confidence intervals, tests of significance (Kish and Frankel 1974).

Design effects are defined as deft^2 = actual variance/simple random variance of same n, both estimated. And values of deft > 1 have been shown for sampling errors not only of means, but also for analytical statistics like differences of means (and Chi square tests), regression coefficients *etc.* It is true that considerable reductions and differences of deft values have been found for some analytical statistics. The differing deft values are not mere necessary mathematical consequences of the sample design, which may be deduced once for all. They have empirical content and therefore they need to be replicated with empirical investigations (Kish and Frankel 1974; Kish 1987, 7.1; Kish 1965, 14.1-14.2; Rao and Wu 1985; Scott and Holt 1982; Skinner, Holt, and Smith 1989). In this paper we investigate the possible effects and the magnitudes of design effects for a set of related statistics that have not been investigated before. On the contrary, in several statistical papers the absence of design effects was merely assumed by the authors (all justly famous), and apparently passed on by the journal referees, without warning the readers. We shall see if deft is reduced or eliminated for this set of analytical statistics (Cochran 1950; Mosteller 1952; Scott and Seber 1983; Seber and Wild 1993).

Furthermore, we also propose explicitly, as has been implied before, that the existence of considerable values of deft is strong evidence for the need for probability selections. It would be difficult to assume a model of a population distribution where the selection design was unimportant (or uninformative) but produced considerable design effects. The reverse does not hold: absence of design effects is necessary but not sufficient evidence for license to neglect probability selection. This proposition gives added importance to our study, which relates deft$(p_i - p_j)$ for analytical statistics to deft(p_i) and deft(p_j) for two of several categories of the same variable.

Section 2 describes the five related problems (designs) for which sampling errors are described in Section 3. Section 4 discusses the empirical evidence in the tables. Section 5 places our findings in the context of earlier work on defts for subclasses and their differences.

[1] Leslie Kish, ISR, University of Michigan, Ann Arbor MI 48106, U.S.A.; Martin R. Frankel, NORC and City University of New York; Vijay Verma, University of Essex, Colchester, C04 3SQ, U.K.; Niko Kaćiroti, Institute of Statistics, Tirana, Albania.

2. SIMILAR STATISTICS FOR FIVE DESIGNS

It has been shown that five designs (problems), of two distinct types, can be treated with the same simple statistics (Kish 1965, Section 12.10). For our empirical and simple presentation we use symbols for sample values (like deft, p_i and n_i), even when occasionally capitals for population values would be more appropriate.

The difference of proportions $p_2 - p_0 = n_2/n - n_0/n$ expresses the desired estimate, where $n = n_0 + n_1 + n_2 + \ldots n_k$ is the sample size, with n units selected and weighted equally. Furthermore, under simple random sampling assumptions, the variance of $(p_2 - p_0)$ is $(1 - f)[p_2 + p_0 - (p_2 - p_0)^2]/(n - 1)$.

Type A Comparisons

1. The difference between two categories $(n_2 - n_0)/n = (p_2 - p_0)$ of a polytomy can represent preference between two parties among several (k) in voting surveys, or between two brands of automobiles in market research, or two of several attitudes, opinions, behaviors on one variable, *etc.* The other $(k - 2)$ choices are summed into p_1 and disregarded in the difference. (Also treated by Scott and Seber 1983.)

2. Rank values of -1, 0, $+1$ (or 0, 1, 2 or c, $c + 1$, $c + 2$) can be assigned to an ordered trichotomous variable without a metric, and viewed as a simple form of the difference of two categories. This form is particularly useful for computations of sampling errors, because all the five designs can use -1, 0, $+1$ for instance as a transformed computing variable.

3. The difference of proportions from two different variables (x and y) may be treated as in (1) and (2). Define as positive in x (or success) only those elements that are positive in x but not in y, so that $n_{10} = n(x_1, y_o)$. Similarly define as positive y the $n_{01} = n(x_0, y_1)$. Then $(n_{10} - n_{01})/n = (p_x - p_y)$ is the net difference in the proportion of positives in x and y. Those that are positives or negatives in both x and y do not count in the differences. Thus we have a case of three categories as in (1) and (2). An example is the difference between the proportions who would "stop all nuclear testing," and those who "want complete nuclear disarmament"; or who would "force Iraq to leave Kuwait" and who would "remove Saddam from power," (Wild and Seber 1993). However, the two categories may also come from two *different surveys* of the same n cases, as in a quality check, or from dual frame observations, or from two waves of a sample. These situations resemble those of (4) and (5).

Type B Comparisons

4. Test-retest and before-after are terms for designs in which the same subjects undergo two observations. Then dichotomous answers $n_2 = n_{10}$ denote the number of negative changes; $n_0 = n_{01}$ the number of positive changes; and $n_{11} + n_{00}$ the sum of the unchanged positives and negatives. Positive and negative answers are respectively denoted here as 1 and 0, and the first and second wave by the order of the subscript. The difference $(n_{10} + n_{11}) - (n_{01} + n_{11}) = n_{10} - n_{01} = n_2 - n_0$ measures the change between positives for the two observations; and $p_2 - p_0 = n_2/n - n_0/n$ measures the change in proportions. (McNemar 1949; Cochran 1950; Mosteller 1952).

5. Matched pairs of n pairs of subjects can also be treated as a generalization of the test-retest design (Mosteller 1952). For example n pairs of randomized subjects may represent experimental versus control treatments; or n pairs of boys versus girls matched on control variables. The statistical treatment $(p_{10} - p_{01})$ of the n pairs of matched subjects is the same as for the n pairs of treatments on the same n subjects (4).

The similarity of statistical treatment for these five designs of two distinct types is convenient, and we present empirical results for both types. "It also has heuristic value that has been overlooked in recent publications (Scott and Seber 1983 and Wild and Seber 1993). The Chi-square test for types 4 and 5 was published early (McNemar 1949; Cochran 1950; Mosteller 1952), and the similarity to the categorical cases 1, 2, 3 was shown" (Kish 1965, 12.10). (Kish was wrong in denoting "trichotomies and matched dichotomies," as "Trinomials and Matched Binomials," which terms refer to IID samples only.)

All of these deal with differences of proportions p_i based on count variables n_i. Extensions to correlated differences $(y_i - y_j)$ for other variables are possible, but not within the scope of our study. Practical examples would include the difference in dollar shares (not only numbers n_i) between two automobile makes from a total of $\sum y_i$ sales.

3. SAMPLING ERRORS AND DESIGN EFFECTS

For simple random samples of size n it can be easily shown (Kish 1965, 12.10) that

$$\text{var}(p_2 - p_0) = \left[\frac{(1 - f)n}{(n - 1)}\right][p_2 + p_0 - (p_2 - p_0)^2]/n.$$

Most of the examples found and shown come from large survey samples, where the $(1 - f)$ can be disregarded. It is worth noting that for the element variance

$$p_2 + p_0 - (p_2 - p_0)^2 = p_2q_2 + p_0q_0 + 2p_2p_0,$$

where the last term $\text{cov}(p_2,p_0) = -p_2p_0$ represents the covariance arising because p_2 and p_0 are competitive parts of the same sample, rather than proportions from independent samples. The difference of proportions squared $(p_2 - p_0)^2$ will usually be a small correction term, and without it we have the equivalent of the variance $(p_2 + p_0)/n$ of two independent Poisson samples. Furthermore, note that (Kish 1965, 12.10):

> The Chi-square test has been applied to some of these problems, treated separately (Cochran 1950; Mosteller 1952; McNemar 1962, p. 225). This is essentially $(n_2 - n_0)^2/(n_2 + n_0)$ the square of the difference divided by its variance, under the null hypothesis $n_2 = n_0$. It applies the exact theories available for tests of null hypotheses in small samples, including the "Yates correction," all based on the assumption of simple random sampling. However, there are great advantages in treating these problems in large samples as estimated means with proper standard errors. First, instead of being confined to testing null hypotheses, we can make inferences with the probability intervals $(p_2 - p_0) \pm t_p\, se\,(p_2 - p_0)$. Second, the formulas for standard errors of complex samples can be applied directly to the mean $(p_2 - p_0)$. Third, the logical structure of this statistic $(p_2 - p_0)$ can be seen more clearly in its application to several distinct problems.

Correlated proportions originate usually in data from complex surveys, and the computations of variance should be appropriate to the sample design. The variance formulas for stratified complex samples can be adopted, but the direct formula has eight terms (Kish 1965, 12.10.3). Instead, it is convenient to translate the problem into a trichotomous variable, with values of -1, 0, $+1$ as in design 2 of Section 2; and the computations of Section 4 used that translation.

Then comparisons between variables and between samples can be facilitated by recourse to the design effects:

$$\text{deft}^2(p_2 - p_0) = \frac{\text{computed variance of } (p_2 - p_0)}{[p_2 + p_0 - (p_2 - p_0)^2]/n}.$$

A few words are needed about limitations on the use of deft as a tool for robust approximations. They serve well for clustered and multi-stage samples using ultimate clusters (primary selections) for computing sampling errors. However, we avoided the problem of weighted samples, because their treatment would be too specific and perhaps too complex. Weighting for nonresponse would not be important for the ratio of $\text{deft}(p_i - p_j)$ to $\text{deft}(p_i)$. However weights for gross inequalities of selection probabilities need specific treatments. Nevertheless, inference and experience indicate that deft values are less affected by weights than are the variances and means themselves. Furthermore we conjecture that the relations we found between the values of $\text{deft}(p_i - p_j)$ and $\text{deft}(p_i)$ will hold also for weighted data, if these are not extreme or pathological.

An approximate but dependable relation of $\text{deft}(p_i - p_j)$ to $\text{deft}(p_i)$ and $\text{deft}(p_j)$ would be useful to allow inferences from the latter, which are routinely and easily computed, to the former that are not. Several alternative conjectures may seem reasonable, and none can be mathematically derived, nor excluded.

1. $\text{Deft}(p_i - p_j) = 1$ of no design effect was assumed implicitly in the five publications referenced in Section 1.
2. $\text{Deft}(p_i) > \text{deft}(p_i - p_j) > 1$ denotes persisting but lower effects than for the $\text{deft}(p_i)$ for proportions. This happens for "crossclasses" and their comparisons (Kish 1987, 7.1). This also seemed reasonable to several experienced statisticians we polled.
3. $\text{Deft}(p_i - p_j) = [\text{deft}(p_i) + \text{deft}(p_j)]/2$ is what we actually found to be a good approximation for all of our data, from different populations and designs. This conjecture seems reasonable, because design effects due to clustering for individual p_i can apply similarly to the variable created from the difference $(p_i - p_j)$ of two of them.
4. Inconsistent results would have been possible, but annoying by preventing inference.

4. EMPIRICAL RESULTS FOR Deft$(P_i - P_j)$

Without strong theoretical or mathematical basis for favoring any of the four alternative conjectures, empirical results about $\text{deft}(p_i - p_j)$ become essential, linking these to the computed values for $\text{deft}(p_i)$. These resemble our more familiar conjectures about $\text{deft}(p_i) = \sqrt{1 + roh[\bar{b} - 1]}$; their value depends on several factors that affect *roh*, the coefficient of intraclass correlation, in addition to the average cluster size $\bar{b}$ (Kish 1965, 5.4, 8.2). The values of $\text{deft}(p_i)$ vary greatly between surveys, also between variables for the same survey (Kish, Groves and Krotki 1976; Verma, Scott and O'Muircheartaigh 1980; Verma and Lê 1995). However, survey statisticians gain knowledge from empirical investigations of sampling errors from diverse surveys, which also permit relating the deft values of complex statistics to the simpler $\text{deft}(p_i)$ (Kish L. 1995; Rao and Wu 1985; Rao and Scott 1987). Similarly, to learn about the relation of $\text{deft}(p_i - p_j)$ to $\text{deft}(p_i)$ we have here empirical results from many variables and from many surveys.

In this first essay into this field we present data from fourteen surveys, which represent a great variety of situations. Eleven surveys presented as 5 sets of results (Figures 1 and 2 and Tables 1-3) deal with paired differences of categories from single surveys (Type A). Three sets of results (Tables 1-3) come from social surveys, followed by two sets (Figures 1 and 2) from the Demographic and Health Surveys on population data. Finally three other sets, each dealing with two waves of data, each based on two reinterviews with the same respondents (Tables 4, 5 and 6), represent type B designs of comparisons.

Tables:

1. The National Election Study of 1986 of the Institute for Social Research of the University of Michigan, $n = 2{,}135$.
2. The National Education Longitudinal Study (NELS) of 1988, the National Opinion Research Center of the University of Chicago, $n = 24{,}355$.
3. The National Longitudinal Study of Labor Market Experience of Youth, conducted by the National Opinion Research Center of the University of Chicago, $n = 5{,}857$.
4. National Election Studies Panels 1990 and 1992, Survey Research Center, Institute for Social Research, Ann Arbor, MI 48106.
5. Panel Study of Income Dynamics 1983 and 1987, Survey Research Center.
6. Americans' Changing Lives 1986 and 1989, Survey Research Center.

Figures:

1. Demographic and Health Surveys of Morocco, Niger, and Colombia, MACRO International.
2. Population Census of Indonesia, Rural Java strata (unpublished data).

We note the following important, useful, EMPIRICAL results.

1) First and foremost: The design effects $\text{deft}(p_i - p_j)$ for the differences are usually NO LESS than the $\text{deft}(p_i)$ for the proportions themselves, and $\text{deft}(p_i - p_j) \simeq 0.5\,[\text{deft}(p_i) + \text{deft}(p_j)]$ approximately in all cases. They vary together, along with the considerable variation for deft values between variables, and also with the lesser variation between pairs of categories for the same variables. Researchers who neglect deft commit the usual under-statement of sampling errors for statistics from clustered surveys. This observation is not only interesting but also a useful model for inference, because the other three sources of variation - across variables, categories within variables, and sampling errors of individual statistics - are all greater.

2) We can find these results in all the 14 sets of survey data in the tables and graphs, and we can illustrate them now with Table 1. Note that defts vary from essentially 1.00 for variable D (problems in country) to as high as 2.32 in variable A (religion) which implies $\text{deft}^2 = 2.32^2 = 5.38$. That our empirical rule (1) holds over the range is reassuring. Such variation between variables in the same sample are common and should force us to abandon the practice of using a common average for all defts of a sample (Verma and Lê 1995; Kish 1995).

   Furthermore, we emphasize here the great variation in deft values for the five categories of the same variable from 1.21 to 2.32 (No. 3 for "fundamental" protestants). It follows that $\text{deft}(p_i - p_j)$ is large only when i or j is category 3 for this variable. These variations among the defts for categories of the same variable mean that they should be computed for all categories rather than for only a single "representative" category. These large possible variations between categories of the same variable are an important new finding in our results, that seems to have escaped notice before.

3) There are also sampling errors in the computed values of the defts. Only statisticians who have computed many sampling errors and design effects seem to get the "feel" for how great these can be. They may be mostly responsible for the few cases where $\text{deft}(p_i - p_j)$ fails to fall between $\text{deft}(p_i)$ and $\text{deft}(p_j)$ and either $\text{deft}(p_i) < \text{deft}(p_i - p_j) > \text{deft}(p_j)$ or $\text{deft}(p_i) > \text{deft}(p_i - p_j) < \text{deft}(p_j)$. Incidentally, these cases also show that our results are not mathematical consequences, but empirically based.

The empirical results presented in Figures 1 and 2 further confirm the findings already presented in Tables 1, 2, and 3. Here also we see that: 1) $\text{deft}(p_i - p_j) \cong [\text{deft}(p_i) + \text{deft}(p_j)]/2$ approximately, along the 45° line; that 2) those equalities hold along a wide range of designs effects; and that 3) the variation between variables is large indeed. This large variation is particularly evident for rural Indonesia, with deft values over 4, hence deft^2 values over 16. These large clustering effects are due to the large cluster sizes: with $\bar{b} = 133$ and 137, the values of $roh = 0.12$ are enough for large defts. Note that these empirical results come both from very diverse populations and diverse variables; different from each other and from the data of Tables 1, 2, and 3. Figure 1 has data from 3 countries (Morocco, Niger and Colombia) hence 6 populations, because the urban and rural defts are quite different. Figure 2 shows results for males and females who are quite distinct populations for the occupational variables, though less so for the educational classes.

The empirical data in the tables of studies 4, 5, and 6 were awaited with doubt and anxiety. True that the preceding five sets resulted in similar conclusions, although they dealt with eleven different populations and scores and variables. But studies 1 to 5 all dealt with pairs of categories from polytomies, designs 1 and 2 of Type A. But now we

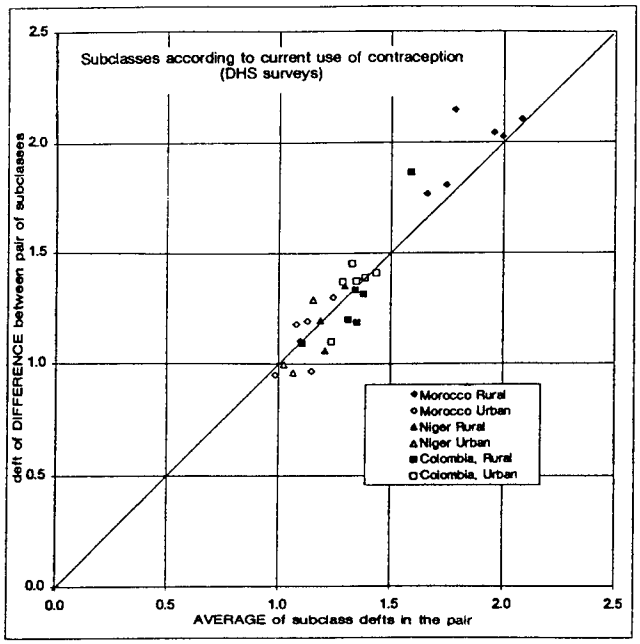

Figure 1. Comparison of deft $(p_i - p_j)$ to the average of deft(p_i),deft(p_j) for categories by current use of contraception*. Illustration of six populations from Demographic and Health Surveys.

* 1 = not using any method of contraception
2 = using only traditional method
3 = using a modern 'reversible' method
4 = sterilised.

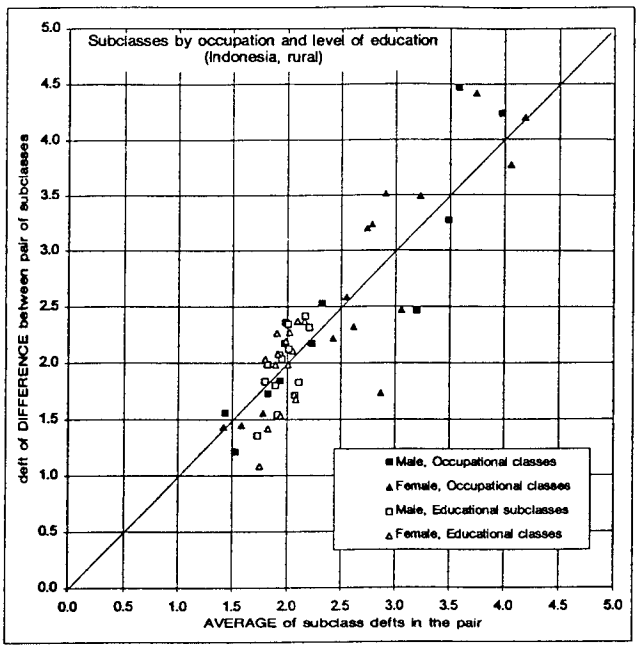

Figure 2. Comparison of deft $(p_i - p_j)$ to the average of deft(p_i),deft(p_j) for categories by occupation and level of education by sex. Illustration from a population census.

Table 1

The National Election Study of 1986 of the I.S.R. of the University of Michigan (n = 2,135)

Categories	Defts for			
$i - j$	P_i	P_j	Average	$(P_i - P_j)$
	A. Religion			
1-2	1.21	1.42	1.32	1.10
1-3	1.21	2.32	1.77	2.02
1-4	1.21	1.50	1.36	1.18
1-5	1.21	1.18	1.19	1.17
2-3	1.42	2.32	1.87	1.93
2-4	1.42	1.50	1.46	1.57
2-5	1.42	1.18	1.30	1.27
3-4	2.32	1.50	1.91	2.03
3-5	2.32	1.18	1.75	2.04
4-5	1.50	1.18	1.34	1.19
Mean	*1.56*	*1.53*	*1.54*	*1.55*
	C. Support Reagan			
1-2	1.32	1.10	1.21	1.07
1-3	1.32	.86	1.09	1.26
1-4	1.32	1.48	1.40	1.50
2-3	1.10	.86	.98	.96
2-4	1.10	1.48	1.29	1.38
3-4	.86	1.48	1.17	1.09
Mean	*1.17*	*1.21*	*1.19*	*1.21*
Overall mean	*1.23*	*1.24*	*1.23*	*1.24*

Categories	Defts for			
$i - j$	P_i	P_j	Average	$(P_i - P_j)$
	B. Abortions Beliefs			
1-2	1.27	.97	1.12	.97
1-3	1.27	1.28	1.28	1.32
1-4	1.27	1.31	1.29	1.36
2-3	.97	1.28	1.12	1.08
2-4	.97	1.31	1.14	1.16
3-4	1.28	1.31	1.30	1.32
Mean	*1.17*	*1.24*	*1.21*	*1.20*
	D. Problems in Country			
1-2	1.07	.94	1.00	.98
1-3	1.07	1.04	1.05	1.09
1-4	1.07	.93	1.00	1.12
2-3	.94	1.04	.99	1.01
2-4	.94	.93	.93	.85
3-4	1.04	.93	.98	.82
Mean	*1.02*	*.97*	*.99*	*.98*

Table 2

The National Education Longitudinal Study (NELS) of 1988, the National Opinion Research Center of the University of Chicago, (n = 24,355)

Categories	Defts for				Categories	Defts for			
$i - j$	P_i	P_j	Average	$(P_i - P_j)$	$i - j$	P_i	P_j	Average	$(P_i - P_j)$
A. Education Status					**B. Classes are boring**				
1-2	1.38	1.22	1.30	1.11	1-2	.99	1.11	1.05	1.04
1-3	1.38	1.14	1.26	1.16	1-3	.99	1.12	1.06	1.07
1-4	1.38	1.19	1.29	1.30	2-3	1.11	1.12	1.12	1.13
1-5	1.38	1.42	1.40	1.54	*Mean*	*1.03*	*1.12*	*1.08*	*1.08*
2-3	1.22	1.14	1.18	1.11	**C. Freedom to pursue interest**				
2-4	1.22	1.19	1.21	1.24	1-2	1.28	1.10	1.19	1.21
2-5	1.22	1.42	1.32	1.45	1-3	1.28	1.08	1.18	1.28
3-4	1.14	1.19	1.17	1.18	2-3	1.10	1.08	1.09	.97
3-5	1.14	1.42	1.28	1.37	*Mean*	*1.22*	*1.09*	*1.15*	*1.15*
4-5	1.19	1.42	1.31	1.20	**D. School offers good jobs**				
Mean	*1.27*	*1.28*	*1.27*	*1.25*	1-2	1.24	1.07	1.16	1.17
E. Religion					1-3	1.24	1.11	1.18	1.24
1-2	2.48	2.83	2.65	2.74	2-3	1.07	1.11	1.09	1.01
1-3	2.48	2.02	2.25	2.09	*Mean*	*1.18*	*1.10*	*1.14*	*1.14*
2-3	2.83	2.02	2.42	2.59	**F. Dad education**				
Mean	*2.60*	*2.29*	*2.44*	*2.47*	1-2	1.61	1.76	1.69	1.83
I. Feel good about self					1-3	1.61	1.68	1.65	1.65
1-2	1.42	1.28	1.35	1.37	2-3	1.76	1.68	1.72	2.48
					Mean	*1.65*	*1.71*	*1.69*	*1.99*
Overall mean						*1.48*	*1.41*	*1.45*	*1.49*

Table 3

The National Longitudinal of Labor Market Experience of Youth, Conducted by the National Opinion Research Center of the University of Chicago, (n = 5,857)

Categories	Defts for				Categories	Defts for			
$i - j$	P_i	P_j	Average	$(P_i - P_j)$	$i - j$	P_i	P_j	Average	$(P_i - P_j)$
A. Chance is important in my life					**B. Something stops me**				
1-2	1.26	1.20	1.23	1.06	1-2	1.07	1.22	1.14	1.04
1-3	1.26	1.18	1.22	1.30	1-3	1.07	1.12	1.10	1.14
1-4	1.26	1.16	1.21	1.28	1-4	1.07	1.09	1.08	1.09
2-3	1.20	1.18	1.19	1.22	2-3	1.22	1.12	1.17	1.28
2-4	1.20	1.16	1.18	1.25	2-4	1.22	1.09	1.16	1.14
3-4	1.18	1.16	1.17	1.05	3-4	1.12	1.09	1.11	1.07
Mean	*1.23*	*1.17*	*1.20*	*1.19*	*Mean*	*1.13*	*1.12*	*1.13*	*1.13*
C. Have control of my life					**D. I am as worthy as others**				
1-2	1.13	1.06	1.10	1.09	1-2	1.17	1.13	1.15	1.16
1-3	1.13	1.10	1.12	1.13	1-3	1.17	1.07	1.12	1.16
2-3	1.06	1.10	1.08	1.07	2-3	1.13	1.07	1.10	1.08
Mean	*1.11*	*1.09*	*1.10*	*1.10*	*Mean*	*1.16*	*1.09*	*1.12*	*1.13*
E. Plans hardly work out					**F. I am satisfied**				
1-2	1.19	1.07	1.13	1.12	1-2	1.19	1.12	1.16	1.16
1-3	1.19	1.13	1.16	1.20	1-3	1.19	1.13	1.16	1.20
2-3	1.07	1.13	1.10	1.08	2-3	1.12	1.13	1.13	1.09
Mean	*1.15*	*1.11*	*1.13*	*1.13*	*Mean*	*1.17*	*1.13*	*1.15*	*1.15*
I. Mother's work									
1-2	1.49	1.36	1.43	1.41					
1-3	1.49	1.52	1.51	1.53					
2-3	1.36	1.52	1.44	1.44					
Mean	*1.45*	*1.47*	*1.46*	*1.47*					
Overall mean	*1.20*	*1.17*	*1.18*	*1.19*					

Table 4

National Election Studies Panels 1990 and 1992, Survey Research Center, Institute for Social Research, Ann Arbor

Categories before/after (90/92)	Defts for			
	P_i	P_j	Average	$(P_i - P_j)$
Strongly approve Bush	1.14	.93	1.04	1.02
Approve Bush foreign policy	.92	1.05	.99	1.00
Strongly disapprove Bush foreign policy	1.23	1.24	1.24	1.32
Approve Bush economy	.97	.94	.96	.96
Strongly approve Bush economy	1.14	1.04	1.09	1.10
Approve Bush	1.00	1.00	1.00	1.00
Strongly disapprove Bush	1.16	1.10	1.13	1.12
Watch campaign on TV	.89	1.55	1.22	1.40
Mean	*1.06*	*1.11*	*1.08*	*1.11*

Table 5

Panel Study of Income Dynamics, 1983 and 1987, Survey Research Center, Ann Arbor

Categories* before/after (83/87)	Defts for			
	P_i	P_j	Average	$(P_i - P_j)$
Live in South	1.22	1.23	1.23	1.11
Age of head of family	1.28	1.33	1.31	1.37
Family size	1.29	1.43	1.36	1.47
Number of children in family	1.23	1.43	1.33	1.49
Work hours of head	1.12	.84	.98	1.03
Age of youngest child	.93	.91	.92	.87
Mean	*1.18*	*1.20*	*1.19*	*1.22*

* All variables are categorized in two categories.

sought data for Type B comparisons from panel surveys, so that we could investigate the conjectures for the test/retest and before/after experimental designs. Mathematically these can be easily shown to resemble polytomies (*i.e.*, tetratomies), but from that to the empirical values of design effects leads through a "black box." Hence these empirical values are so much more valuable and remarkable. Here we found considerable design effects for Chi square tests for analytical comparisons.

5. PRESENT FINDINGS IN THE CONTEXT OF RELATED RESEARCH

A great deal of empirical information is available from previous work by the authors and by others on design effects for the total sample, for subclasses, and for differences, for diverse variables and designs. It would be useful to put the present findings in the context of that work.

It has been found that nature of the survey variables being estimated is a major (often the main) determinant of the magnitude of the design effects: vastly differing defts can occur for different types of variables even with the same samples or with similar designs. For this reason we have always recommended that defts be computed for many different variables, while it is generally less important to compute them for many different subclasses, especially for different categories of subclasses defined in terms of the same characteristic.

The present findings illustrate that defts can differ greatly also among different categories of the same survey variable, estimated with the total sample as the common base. Therefore each individual category and each difference between pairs of categories, even when defined in

Table 6

Americans' Changing Lives, 1986 and 1989, Survey Research Center, Ann Arbor

Categories before/after	Defts for				Categories before/after	Defts for			
	P_i	P_j	Average	$(P_i - P_j)$		P_i	P_j	Average	$(P_i - P_j)$
A. Get together with friends					**B. How often do you exercise**				
Once a week	1.30	1.26	1.28	1.28	Often	1.51	1.67	1.59	1.26
2-3 a month	.88	1.00	.94	1.02	Never	1.62	1.97	1.80	1.41
Mean	*1.09*	*1.13*	*1.11*	*1.15*	*Mean*	*1.56*	*1.82*	*1.70*	*1.34*
C. How Satisfy Are You					**D. How do you like your home**				
Very satisfy	1.28	1.21	1.25	1.33	Very much	1.24	.90	1.07	.91
Not satisfy	1.04	1.16	1.10	1.00	Not much	1.33	.98	1.16	1.12
Mean	*1.16*	*1.19*	*1.18*	*1.17*	*Mean*	*1.29*	*.94*	*1.12*	*1.02*
E. How often work in garden					**F. I have a positive attitude**				
Often	1.40	1.16	1.28	1.19	Agree	1.10	1.33	1.22	1.19
Rarely	.91	1.11	1.01	1.18	Disagree	1.05	1.28	1.17	1.21
Never	1.66	1.17	1.42	1.26	*Mean*	*1.08*	*1.31*	*1.20*	*1.20*
Mean	*1.32*	*1.15*	*1.24*	*1.21*					
Overall mean	*1.25*	*1.26*	*1.26*	*1.18*					

terms of the same survey variable, needs to be regarded, in a sense, as a separate variable in its own right for the purpose of computing and analyzing design effects.

As to the relationship between defts for subclasses and subclass differences, previous research has mostly dealt with the following situation. With the total sample n partitioned into subclasses i of size $n_i = p_i.n$, deft(r_i) values for statistics r_i (such as a proportion m_i/n_i, mean $\sum y_i/n_i$, ratio $\sum y_i/\sum x_i$), estimated over subclass elements n_i as the base, are related to deft(r) for the same variable estimated with the total sample as the base. Similarly, deft$(r_i - r_j)$ for subclass differences are related to deft(r_i),deft(r_j) based on individual subclasses and to deft(r) based on the total sample. Numerous computations confirm these relationships to be in accord with our conjecture (2) of section 3:

$$\text{deft}(r) > \text{deft}(r_i); \quad \text{and} \quad \text{deft}(r_i) > \text{deft}(r_i - r_j) > 1.$$

These effects of covariances on design effects of clustered samples are essentially empirical (even sociological in a broad sense); and they must be so verified.

Similarly with our newly discovered relationship for $(p_i - p_j)$ for two categories, which are so different from the above. The relations $\text{deft}(p_i - p_j) \cong [\text{deft}(p_i) + \text{deft}(p_j)]/2$ are also empirical and approximate and they must be verified over and over again. But they seem to be widely applicable in our data, and clearly better than the other assumptions, such as $\text{deft}(p_i - p_j) = 1$ that have been often assumed until now.

ACKNOWLEDGEMENTS

The authors wish to thank the editor and referees whose suggestions made this paper both shorter and better.

REFERENCES

COCHRAN, W.G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256-66.

DEMING, W.E. (1953). On the distinction between enumerative and analytic studies. *Journal of the American Statistical Association*, 48, 244-45.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

KISH, L. (1987). *Statistical Research Design*. New York: John Wiley and Sons.

KISH, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.

KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, (B), 36, 1-37.

KISH, L., GROVES, R.M., and KROTKI, K. (1976). *Sampling Errors for Fertility Surveys*. Occasional Paper No. 17, *World Fertility Surveys*. International Statistical Institute: The Hague.

LÊ, T., and VERMA, V. (1995). *Sample Designs and Sampling Errors for the DHS*. Calverton MD: MACRO International.

McNEMAR, Q. (1949). *Psychological Statistics*. New York: John Wiley and Sons.

MOSTELLER, F. (1952). Some statistical problems in measuring the subjective responses to drugs. *Biometrika*, 8, 220-226.

RAO, J.N.K., and SCOTT, A.J. (1987). On simple adjustments to Chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.

RAO, J.N.K., and WU, C.F.S. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

SCOTT, A.J., and HOLT, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-54.

SCOTT, A.J., and SEBER, G.A.F. (1983). Difference of proportions from the same survey. *The American Statistician*, 37, 319-20.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley and Sons.

VERMA, V., and LÊ, T. (1995). Sampling errors for the DHS survey. 50th Session of the International Statistical Institute, Beijing.

VERMA, V., SCOTT, C., and O'MUIRCHEARTAIGH, C. (1980). Sample designs and sampling errors for the World Fertility Surveys. *Journal of the Royal Statistical Society (A)*, 143, 431-473.

WILD, C.J., and SEBER, G.A.F. (1993). Comparing two proportions for the same survey. *The American Statistician*, 47, 178-181.