Appendix 1

RELIABILITY

SUMMARY Reliability is technically not a form of validity. The reliability of the instrument was tested for all pairwise combinations for the years 1994 through 1999 (15 individual contrasts) using the Kruskal-Wallis test. In all cases, the findings confirmed reliability by showing that differences in the data between years could not be established. The survey instrument is judged reliable by the accepted standards of the discipline.

"In research, the term reliability means 'repeatability' or 'consistency'."(Trochim, 1999e) The definition of reliability implies that it has two distinct components. Traditional reliability measures seek to establish reliability based on multiple scorings of individual respondents (i.e., repeatability component of Trochim's definition). This is appropriate since many of these tests were applied in the validation of constructs that were hypothesized to be fixed components of a human being (e.g., "introversion").

Organizational Engineering theory, however, posits that people operate within ranges on the underlying method and mode scales—a "built-in" source of variation that is imbedded in the theory. In addition, the theory postulates that people are responsive to their environment. It proposes that people will change their strategic approach (i.e., their method/mode range election) in response to personal environmental changes (Salton, 2000, pp. 53-59). Since reliability can be considered an element of construct validity (Moss,

1994), it would be an error to apply the traditional repeatability tests of reliability. To do so would violate and invalidate the underlying construct that the test attempts to validate. In other words, procedures such as Test-Retest reliability cannot be used without undermining the very foundation of the validity study itself.

Consistency is the second component of the definition of reliability. "We judge the reliability of the instrument by estimating how well the items that reflect the same construct yield similar results." (Trochim, 1999f) Applying the same measure to different subjects and obtaining the same expected result can thus be interpreted as evidence of consistency.

One method of determining this type of consistency is a redundancy strategy. Here the respondent is repeatedly asked the same question at multiple points in an instrument. The responses can then be compared and the degree of correlation viewed as an index of consistency within the instrument. This method is typically employed to ferret out deceptive responses.

The redundancy method is not applicable to the Organizational Engineering survey instrument. The respondent is not answering a simple question. Rather, he or she is expressing a preference for one response versus three other alternatives.

For example, in one place the respondent is presented a selection of "I respond fast" and in another "I react fast." However, each of these selections is set off against different optional alternatives generated by the other potential elections on the method and mode dimensions. Thus it is not inconsistent for a respondent to elect the "I respond fast" option in one case and reject "I react fast" in another. It is merely a statement of preferences among the alternatives provided. Thus the specification of the instrument precludes the use of the traditional consistency measures based on redundancy. If applied, they are likely to yield a false negative, since they presume that the same thing is being asked multiple times.

Parallel forms reliability is an accepted strategy in the social sciences and is used to test the consistency and repeatability of the instrument simultaneously. The parallel forms strategy typically involves applying two instruments purporting to measure the same things to the

same population and comparing the results (Trochim 1999f). A variation the parallel forms methodology can be obtained by applying the same instrument to various populations. Highly correlated results could imply the existence of an underlying consistency sufficient to validate the reliability of the instrument.

In this study, the database was segregated into people who had been administered the survey in the years 1994, 1995, 1996, 1997, 1998, and 1999, creating six separate populations. The variation of the parallel forms test considers that these subsets are "samples" of a larger underlying population. If the underlying population suffered no major environmental changes, it would be expected that the underlying strategic preferences would remain constant.

The years 1994 through 1999 are similar in terms of their macroeconomic and social conditions. Thus, without large-scale dislocations, it is expected that the average personal environment of the underlying population was constant over this time period. Therefore, a testable hypothesis based on the consistency component of the reliability criteria can be stated as:

Null: The strategic postures of the population for years 1994, 1995, 1996, 1997, 1998, and 1999 are statistically indistinguishable.

A failure to reject the null hypothesis would serve as evidence of the reliability of the survey instrument. The instrument would have yielded consistent results over a long time period. The use of six years greatly strengthens the assertion of validity since this range of years provides fifteen opportunities for rejection (1994 vs 1995, 1994, vs 1996, etc.).

The choice of the variable to represent the strategic posture is informed by the underlying theory. "Strategic patterns are most useful in characterizing lengthy streams of decisions and overall strategic postures. Strategic styles are generally more useful in predicting transactional characteristics of individual or shorter streams of decisions." (Salton, 2000, p.86)

Since the hypothesis seeks to test whether the overall postures of the population remain stable over years, the single most appropriate representation of the strategic posture for purposes of testing the reliability hypothesis is the dominant strategic pattern—a combination of the person's primary and secondary strategic style.

The choice of the test statistic to employ in the validation is governed by the character of the data being addressed. For each year, a hypothesis that the strategic pattern characteristics followed a normal distribution was tested using the Shapiro-Wilk test (in the case of the year 1997, Stephens' test was used because the size exceeded the limits of the Shapiro-Wilk test). All six patterns tested resulted in rejection of these null hypotheses at the .01 significance level. In other words, normal distribution of the data could not be assured and statistical tests based upon that normality assumption could not be understood to produce reliable results.

The Kruskal-Wallis test is a nonparametric alternative to one-way ANOVA and is a straightforward generalization of the Mann-Whitney U test for two independent samples. A significance criterion of .05 was chosen for this experiment. The null hypothesis is stated as:

Null: The database populations are drawn from the same underlying population and this population has remained stable.

Table 14

DATABASE NORMALITY TESTS

Year	*N*	Test	Statistic	*p*
1994	158	Shapiro-Wilk	W = .8975	.0001
1995	1082	Shapiro-Wilk	W = .9269	.0001
1996	1891	Shapiro-Wilk	W = .9217	.0001
1997	2453	Stephens	D = .1176	.01
1998	1866	Shapiro-Wilk	W = .9197	.0001
1999	1271	Shapiro-Wilk	W = .9192	.0001

The Kruskal-Wallis procedure requires approximate equality of variance over all groups. Therefore, Levene's test was used to test the hypothesis that all years had equal variance for the measure of pattern. The test obtained Levene's statistic F = 1.113, with a corresponding *p*-value of 0.351. Thus the hypothesis was not rejected, and no significant evidence of different variances was found.

The Kruskal-Wallis test was then applied to the sample population. The overall test obtained a value of the Kruskal-Wallis test statistic $H = 1.809$, with a significance level p of 0.875, failing to reject the overall null hypothesis at the .05 level. No significant evidence was found of differences in the measure of pattern over the different years.

Table 15 contains the results of the multiple comparison of mean ranks across all possible pairs of years. The Tukey-Kramer procedure (Kirk, 1982, pp. 119-120), a well-known a *posteriori* method for evaluating pairwise comparisons, was used to control Type I error. The failure to reject the null hypothesis (that there was no difference in mean rank) in all fifteen of the year comparisons, along with the failure to reject the overall null hypothesis, provides strong evidence of the reliability of the survey instrument. This judgment is strengthened even further when the large number of observations in each pair is considered.

The statistical tests conducted using a database of respondents provide strong evidence that the survey instrument is valid on the dimension of consistency over time.

Table 15

PAIRWISE MEAN RANK COMPARISONS

	Years	q	p	Reject H_0
1.	1994 vs. 1995	0.397	.999	No
2.	1994 vs. 1996	0.751	.999	No
3.	1994 vs. 1997	0.808	.999	No
4.	1994 vs.1998	0.656	.999	No
5.	1994 vs.1999	0.905	.999	No
6.	1995 vs.1996	0.745	.999	No
7.	1995 vs.1997	0.891	.999	No
8.	1995 vs.1998	0.537	.999	No
9.	1995 vs.1999	1.027	.999	No
10.	1996 vs.1997	0.134	.999	No
11.	1996 vs.1998	0.241	.999	No
12.	1996 vs.1999	0.388	.999	No
13.	1997 vs.1998	0.390	.999	No
14.	1997 vs.1999	0.289	.999	No
15.	1998 vs.1999	0.604	.999	No