Journal of Management
1990, Vol. 16, No. 3, 665-674

Positive-Findings Bias in QWL Studies: Rigor and Outcomes in a Large Sample

Robert T. Golembiewski
University of Georgia
Ben-Chu Sun
National Chengchi University

This report seeks to determine whether the high success rates observed in a large survey of QWL evaluative studies (N = 231) can be substantially explained in terms of the lack of rigor of research methodology and design, as the literature critical of QWL often proposes. This study finds statistically significant support for a positive-findings bias hypothesis, but rigor explains less than 7% of the variance in outcomes. This implies only modest support for the position that attractive QWL results can be substantially accounted for by a positive-findings bias.

Much recent attention focuses on assessing the efficacy of attempts at planned change in organizations, especially via Organization Development (OD) and Quality of Working Life (QWL). Much work focuses on OD (e.g., Margulies, Wright, & Scholl, 1977; Morrison, 1978; Golembiewski, Proehl, & Sink, 1981, 1982; Nicholas, 1982; Terpstra, 1982; Mitchell, 1981; Woodman & Wayne, 1985), and that subliterature isolates success rates that are substantial and, indeed, formidable. If anything, QWL applications show an even stronger record of intended outcomes, as over 3 dozen separate surveys indicate (Sun, 1988; Golembiewski & Sun, 1988, 1989).

This article seeks to build on the available work about QWL applications, while transcending it and similar research in OD in three significant particulars. First, existing QWL surveys usually feature a small number of applications, ranging from 30-50 (Sun, 1988: chapter 2). This precludes sensitive analysis, and results in sparsely populated or empty cells even in bivariate comparisons.

Second, survey studies of both QWL and OD applications almost always are narrowly outcome focused. Did applications succeed or fail, and in what proportions? Rare attention gets directed at the rigor of methodology and design of evaluative studies. Moreover, survey studies typically rely on a single measure of rigor (e.g., Terpstra, 1981; Woodman & Wayne, 1985), on those few occasions when that significant issue gets any attention.

Address all correspondence to Robert T. Golembiewski, University of Georgia, Baldwin Hall, Athens, GA 30602.

Consequently, the QWL literature is open to suspicion about the validity of its findings, although that suspicion has nowhere been tested in a large population of evaluative studies. The specific focus here is on the positive-findings bias, which basically proposes that attractive QWL results are artifacts of poor methods and designs. The essence of the notion is that an inverse relationship exists between the rigor of an evaluation and the success of the intervention being evaluated. As rigor increases, this view implies, so will studies reveal fewer attractive effects for QWL, and perhaps no change or even negative effects.

Third, QWL evaluative studies usually deal with "soft" outcomes, such as self-reports about productivity. This leaves results open to charges of being superficial feel-goodisms, of being Hawthorne effects, and so on (Bass, 1983).

Survey studies that overcome one of these three limitations typically fall victim to the others. For example, one OD survey encompasses 532 cases, but differentiates neither degrees of methodological rigor nor studies providing hard data about outcomes from that majority offering only soft data (Golembiewski, Proehl, & Sink, 1981, 1982). Similarly, one OD survey admits only studies providing hard data, but includes only a handful of studies and entirely neglects methodological rigor (Nicholas, 1982).

This article reports on an effort to minimize these three limitations in assessing the efficacy of QWL interventions. It considers unpublished as well as published evaluative studies, while testing whether or not attractive success rates are the unfortunate products of faulty methodology and design. We know of no similar survey of QWL applications that seeks to learn from as well as augment the sparse and inconclusive attention devoted to the positive-findings bias in OD (e.g., Bass, 1983). Some observers report a positive-findings bias in OD (Terpstra, 1981), other studies fail to replicate this finding (Bullock & Svyantek, 1983, 1985), and still others suspect a positive-findings bias in one class of OD interventions, but not in other classes (Woodman & Wayne, 1985).

Method

Five sections outline the present approach to minimizing the three limitations of QWL evaluative studies. These sections, in turn, briefly introduce the pool of QWL studies, detail three estimates of rigor, sketch ways of estimating the success of individual QWL applications, provide estimates of the reliability of scoring, and outline the present analytic approach.

Sample of QWL Studies

A multipronged search sought to develop a comprehensive sample of QWL evaluative studies, both published and unpublished, for the interval 1965-1987. All acceptable studies must provide hard measures of outcomes (e.g., objectively measured output, turnover, absenteeism, cost of raw materials). The search involved:

1. discovery of over 3 dozen bibliographies
2. physical examination of 90 English-language periodicals
3. review of *Dissertation Abstracts*, as well as of the proceedings of several

professional associations (e.g., Academy of Management, American Psychological Association)

4. a mail solicitation of about 100 QWL practitioners

The search isolates 231 studies[1], mostly from published sources but with 13% from unpublished sources like in-house memos or consultant reports. Although hard measures of outcomes are required for inclusion in the sample, all 231 studies also report soft outcomes (e.g., self-reports about satisfaction, cooperation).

Two other features of the sample of QWL evaluative studies deserve note. This sample includes cases from numerous collective sources (e.g., other surveys, bibliographies) and adds to their number by over a third. Moreover, domain issues are not crucial here. Conceptually, most observers define OD broadly enough so that it subsumes QWL as one major class of interventions. In practice, OD and QWL applications share values centering around participation and involvement, but they often differ in a range of particulars: QWL is more likely to deal with unionized employees and OD with management, QWL emphasizes structure and OD has a dominant focus on interaction, and so on (Shelley, 1989). Except in a few cases of egregious mislabelling, the author's description of a case as "QWL" is accepted. More specifically, the present sample encompasses nearly 20 distinct kinds of interventions. They include goal setting, structuring work and social relations in autonomous groups, work redesign and especially job enrichment, Quality Circles, and other applications that combine two or more distinct interventions of the kinds illustrated.

Estimates of Rigor

Three estimates of rigor are employed in order to avoid the common reliance on a single measure. An overview of how these measures assess rigor is helpful:

1. *Terpstra's M/D Score*. This measure of methodology/design rigor is employed in an early analysis of positive-findings bias in OD research (Terpstra, 1981). Six dimensions are assigned a score (1 or 0), depending on whether or not a specific study is considered "rigorous" on each dimension. A study with a score of 6: (a) uses an acceptable sampling strategy such as a full census, (b) has a sample size greater than 30, (c) employs a control group, (d) uses random assignment to treatment or control, (e) provides at least one pre-test and one post-test estimate of effects, and (f) achieves $p < .05$ on statistical tests employed.

2. *Woodman and Wayne's M/D Score*. The developers of this measure propose to improve on Terpstra's version by adding three criteria to his list of six, basically (Woodman & Wayne, 1985). Three additional points are assigned to studies that (g) show no reliabilities $< .60$ and provide some evidence of validity, (h) include hard data relevant to objective criteria, and (i) employ an appropriate multivariate analytic procedure.

3. *Morrison's I/E Validity Score*. She focuses on Campbell and Stanley's (1963) detailed tests of how a research design can eliminate threats to validity. Morrison's (1978) I/E scores range from 0 to 12, with the higher scores indicating greater effectiveness in eliminating threats to validity. Rigorous studies seek

[1]A complete bibliography of these 231 applications is available from Robert T. Golembiewski, University of Georgia, Baldwin Hall, Athens, GA 30602. The complete data set also is available.

to minimize or eliminate these four threats to external validity: interaction of test and treatment, interaction of selection and treatment, reactive features that limit generalizability of results, and multiple-treatment interferences. In addition, a rigorous evaluative study also deals with these eight threats to internal validity:

1. changes due to factors other than the intervention
2. maturation of subjects or groups
3. testing effects
4. instrument effects
5. regression effects
6. selection biases
7. mortality of subjects
8. interaction effects of two or more of the seven threats above.

Although the three measures of "rigor" tap similar dimensions, each appears in the analysis below. Specifically, Terpstra's M/D score and that of Woodman and Wayne correlate .91; and Morrison's I/E Validity score correlates .60 and .67 with the two M/D measures, respectively. These coefficients imply some redundancy in subsequent ANOVA analysis, which is justified by the need for detailed attention to each of the three measures at this early stage of analysis. Given substantially the same patterns of findings for the three measures of rigor, subsequent analyses will have strong justification for using either a composite rigor score or MANOVA.

Estimates of QWL Success Rates

The chief criterion for inclusion in the present pool is that a study assesses QWL outcomes with hard or objective data, but all studies also provide soft data about outcomes. This analysis scores 16 objective outcomes (e.g., quantity of output, various costs such as those for materials and labor, and personnel turnover). Scoring also differentiates 18 self-report outcomes (e.g., job involvement, various facets of satisfaction with work, and organizational commitment).

For each QWL study, this analysis estimates both hard-criteria and global outcomes, with the latter combining objective and self-report data. In both cases, four categories of outcomes are distinguished:

I. Definite balance of highly-positive and intended effects
II. Definite balance of positive and intended effects
III. No appreciable effects
IV. Negative effects

In general, this study proposes that QWL applications will have a consistent set of intended effects. For example, following QWL applications, productivity should increase, employees will report greater satisfaction, and job commitment will grow.

In more detail, "highly positive and intended effects" include statistically significant changes ($p < .05$) or those with a magnitude of 10% or greater where no statistical tests are employed. Category I assignments require that more than half of all pre- versus post-test comparisons meet one or both of these standards, with most other changes also falling in the expected direction. Category II assignments require that most comparisons are in the intended direction and often in-

clude a substantial proportion of statistically significant changes. Small and random changes dominate in Category III, and Category IV includes most applications with more than a sprinkling of unexpected effects and virtually all such cases that attain statistical significance.

Hard-criteria and global outcomes are governed by the same conventions for assignment. Global outcomes are based on all self-report data in a particular evaluation in addition to all objective variables. Hard-criteria outcomes only involve the latter variables.

Inter-Observer Reliabilities

The evidence implies that little variance can be accounted for by differences between coders. Three raters are used, and agreement between pairs of observers ranges from the low- to mid-.90s for the several outcomes and rigor codings. The measure of agreement assigns either a 1 or 0 to each pair of codes, depending upon whether they agree or disagree. The total points assigned are then divided by the total number of pairs, and multiplication by 100 generates a percentage estimate of agreement. There are no missing data for rigor or outcomes.

A conservative convention applies to all codes-in-disagreement, if discussion fails to resolve them *after* the reliabilities are calculated. Each such case is assigned the lower/lowest score in contention. For outcomes, the typical case of disagreement involves a judgment whether a QWL application merits I or II. The final assignment is a II, if discussion does not lead to consensus. Similarly, disagreement about any component of the three rigor scales that remains after discussion results in a 0 code. As noted, each rigor component is scored 1 or 0 depending upon whether or not its treatment is seen as contributing to validity.

In sum, the codings for outcome and rigor involve many judgment calls, and this project reflects a mixed but generally positive record of dealing with them, as three points imply. First, the interrater reliabilities are substantial, as estimated *before* discussion leading to a final judgment. This implies effective training of coders.

Second, conservative conventions reinforced by training govern judgments about both outcomes and the multidimensional scales used to assess rigor. This no doubt helps account for the substantial reliabilities before discussion, but it comes at the cost of introducing a systemic bias that may effect validity. The point applies with special force to Morrison's scale, which requires difficult judgments as to whether particular threats to validity are addressed in any evaluative study. Here, as elsewhere, individual raters are to assign a lower code in cases of any doubt or ambiguity.

Third, this study does a substantial if incomplete job of meeting a comprehensive list of 14 criteria for surveys of this kind (Bullock & Svyantek, 1985: 114-115). A well designed survey

1. Uses a theoretic model
2. Identifies its study domain precisely
3. Includes all publicly available studies
4. Avoids selecting studies in terms of rigor, etc.
5. Publishes or makes available the pool of studies

6. Selects and codes variables on theoretic grounds
7. Provides details about the coding scheme and resolution of problems in its application
8. Uses multiple raters and assesses interrater reliability
9. Reports on all variables analyzed in order to avoid problems with chance relationships in a subset
10. Publishes or makes available the full data set
11. Considers alternative explanations for findings
12. Limits generalization of results to the specified domain
13. Reports study characteristics to shed light on domain analyzed
14. Reports study in sufficient detail to permit direct replication

Reflecting the limitations of the journal article genre, #7 and 14 are the least substantially met criteria in this report.

A fourth issue about coding constitutes a possibly severe constraint on the findings: both rigor and outcome scores are assessed by the same raters. Independence could have been assured, as by using two separate panels of raters, each blind to the study's basic purpose (Woodman & Wayne, 1985). Here, the major defenses against contamination are the guidelines for estimating outcomes and training.

Analytic Procedures to Test for Positive-Findings Bias

In sum, this study employs 12 separate tests of the association of rigor with QWL outcomes. Because rigor is estimated in three ways, and because both hard-criteria as well as global assessments of outcomes are made, this requires 6 tests of the positive-findings bias. In addition, because of the small number of cases rated as having outcomes III and IV, this study also will test for associations of rigor with three categories of outcomes—I, II, and III plus IV. The infrequent III and IV assignments have several possible interpretations, and combining the two categories provides guidance in evaluating those interpretations as well as tests for effects of small subsample size. This useful exercise adds 6 tests for a positive-findings bias.

When one-way analysis of variance (ANOVA) isolates statistically significant variance in rigor and outcomes scores, analysis will be supplemented in two ways. The statistical significance of all possible paired comparisons will be assessed via the Least Significant Difference test, or LSD, as modified for unequal subsample sizes. The direction of all possible paired comparisons also will be assessed for consistency or contrariness with the positive-findings bias.

For this study, evidence supporting a positive-findings bias requires that the higher the rigor of a QWL evaluative study, the poorer its associated outcomes. For each of the 12 tests detailed above, then, *perfect* support for the hypothesis will require that: $p < .05$ for all ANOVA tests; all possible paired comparisons show that rigor is inversely associated with the favorableness of QWL outcomes; and all paired comparisons in the intended direction attain $p < .05$, whereas none in the contrary direction do so.

Results

Review of the results proceeds on two tracks. First, the form of analysis and of the results will be illustrated by focusing on one of 12 individual analyses detailed above. Second, the results of all 12 analyses will be summarized.

An Illustration: Rigor and Hard-Criteria Assessment

Table 1 refers to analysis of three measures of rigor and hard-criteria assessment, this time rated in terms of four categories of outcomes. Four points highlight trends in the data. First, all three *F*-scores in Table 1 indicate non-random variance, but eta^2 indicates that a bit less than 6.7% of the variance is accounted for, on average.

Second, a third of all paired comparisons in Table 1 achieve statistical significance as well as fall in a direction consistent with the positive-findings bias hypothesis, and the direction of an additional 26% of the comparisons is consistent with that hypothesis. Note that 6 paired comparisons are possible for each measure of rigor and the four categories of outcomes. Support for the positive-findings bias requires that for each method the rigor of evaluative studies rated I should be less than for those rated II, I less than III, and so on. Because there are 3 measures of rigor, Table 1 involves 3 X 6, or 18, paired comparisons.

Third, nearly 40% of all paired comparisons fall in a direction contrary to the hypothesis, and a bit over 11% attain statistical significance. The contrary cases all involve outcome category IV, or negative effects. All three measures of rigor tend to rise successively for outcomes I, II, and III, and then fall for outcome IV. For example, Morrison's I/E rigor score for outcome IV is not only the lowest in Table 1, but it is significantly lower than the rigor scores for both outcomes II and III.

Fourth, the pattern for each of the three measures of rigor is quite similar. Only minor variations exist.

Summary: All Measures of Rigor for All Outcomes

The pattern for the illustration above—for four categories of hard-criteria assessments of QWL outcomes—also characterizes the three other sets of associ-

Table 1
Methodological Rigor and Hard-Criteria Outcomes

Outcomes of QWL Interventions	*N* =	Means, Rigor Scores		
		Terpstra's M/D Score	*Woodman and Wayne's M/D Score*	*Morrison's I/E Score*
I. Highly positive and intended effects	160	2.62	3.20	4.40
II. Definite balance of positive and intended effects	50	2.98	4.00	5.12
III. No appreciable effects	13	3.62	5.16	7.62
IV. Negative effects	8	2.75	3.75	2.75
	F =	3.63	6.80	5.81
	p =	.0137	.0002	.0008
	eta^2 =	.05	.08	.07

ations summarized in Table 2A-C. Table 2D summarizes the results discussed in connection with Table 1.

Five points characterize the overall pattern in Table 2. First, nearly 92% of the ANOVAS for overall rigor and outcomes achieve statistical significance. The appropriate tables underlying the summary in Table 2 are not reprinted to conserve space, but one of them shows that the single exception to our first point approaches statistical significance ($F = 2.37, p = .09$)

Second, a noteworthy proportion of the differences between the paired comparisons are quite robust as well as in the direction supporting a positive-findings bias. Specifically, considering those cases attaining statistical significance, 35.2% of all paired comparisons are significantly different as well as directionally consistent with the hypothesis. This record is substantially greater than chance.

Third, over 3 of every 10 paired comparisons fall in the direction consistent with the hypothesis of a positive-findings bias. In 7 of 10 cases, then, the greater the methodological rigor of a QWL evaluation study, the less favorable is the trend in outcomes.

Fourth, outcome/rigor pairs inconsistent with the hypothesis of a positive-findings bias are in a clear minority and, moreover, such differences almost never achieve statistical significance. Specifically, over 29% of all paired comparisons fall in a direction contrary to the hypothesis, and less than 4% achieve the .05 level.

As with Table 1, all of the cases in Table 2 falling in a contrary direction—16, to be exact—involve QWL outcomes rated IV. Although only 2 of those 16 attain statistical significance, support of the positive-findings bias requires that rigor scores be highest for outcome IV.

Fifth, eta^2 averages .067 for the 11 statistically significant ANOVA runs. This is marginally higher than for the illustration in Table 1.

Table 2
Summary, Tests for Positive-Findings Bias

	ANOVA Outcomes and Rigor, $p < .05$	Paired Comparisons			
		In Consistent Direction and Statistically Significant	In Consistent Direction	In Contrary Direction	In Contrary Direction and Statistically Significant
A. Global Outcomes, 4 Outcome Categories	3 of 3	5 of 18	11 of 18	7 of 18	0 of 18
B. Hard Criteria Outcomes, 4 Outcome Categories	3 of 3	6 of 18	11 of 18	7 of 18	2 of 18
C. Global Outcomes, 3 Outcome Categories	3 of 3	5 of 9	7 of 9	2 of 9	0 of 9
D. Hard Criteria Outcomes, 3 Outcome Categories	2 of 3	6 of 9	9 of 9	0 of 9	0 of 9
Totals	11 of 12	19 of 54	38 of 54	16 of 54	2 of 54
Means, %	91.7%	35.2%	70.4%	29.6%	3.7%

Discussion

This analysis permits six major conclusions. These conclusions suggest how this analysis transcends earlier work, how it supports the hypothesis of a positive-findings bias, how that support needs to be interpreted carefully, and how the present analysis can be extended to encompass issues beyond the present scope.

Although not ideal, to begin, the present test of the positive-findings bias has a variety of advantages over earlier research dealing with QWL evaluations. The population of cases is more comprehensive, and can make a stronger claim to representativeness. In addition, three accepted measures of rigor are used, and all generate similar patterns of results. Moreover, hard-criteria effects are required of all cases admitted to analysis, which precludes criticism of the results as based on "mere self-reports." As noted, global outcomes include soft data to provide a check on the hard-criteria assessment.

Consequently, special weight seems appropriate for the present conclusion that the positive-findings bias accounts for only modest variance (6.7%) in QWL outcomes. Prior evidence for OD studies is inconclusive. Terpstra (1981) finds evidence of an inverse relationship between rigor and outcomes in OD, but Bullock and Svyantek (1983) do not, although both studies draw their populations from the same journal during the same interval. Woodman and Wayne (1985) attempt a similar test, and it suggests that a positive-findings bias may exist for some classes of OD interventions but not others.

In addition, this analysis provides perspective on the issue of whether or not "positive results" have a better chance of publication, a common point in contention. Any such effect in this case may be diluted by the long period over which cases in our panel were published, but analysis shows the success rates of published versus unpublished studies differ only randomly ($\chi^2 = 5.58$, $p = 0.13$). To illustrate, nearly 85% of the 26 unpublished cases fall in hard outcome categories I and II. This is lower than for the published cases (91.7%), but not significantly so.

A related finding is that only 8 QWL applications have a Category IV outcome—that is, negative effects. Are these cases somehow similar in other regards? Individual inspection suggests no obvious similarities in date, worksite, intervention, and so on.

So evidence supporting a positive-findings bias in QWL studies not only must be appropriately tethered, but two additional points apply. It is not clear why the positive-findings bias applies least to outcomes rated IV—that is, negative effects. This may reflect attenuated distributions—of methodological rigor or of outcomes, or both. Alternatively, it may simply be that negative outcomes plainly advertise themselves, *whether or not* research designs are rigorous. Perhaps, although this is a bit of a stretch, a kind of incompetence identity exists—poor evaluative designs may occur along with careless or inept interventions, which places *the* issue in the specific implementor/evaluator rather than in QWL.

Finally, the size of the present QWL sample permits perspective on the consistency of these findings. Consider only two possibilities: consistency between the several classes of QWL interventions, and consistency over time.

As for QWL classes, this population encompasses 17 distinct varieties of

QWL interventions aggregated into four broad classes. Consistency of the present findings *within* each of these four broad categories will provide a powerful test of this analysis. The small populations underlying most other surveys of planned change preclude such a test, although some observers suspect that differences exist (e.g., Terpstra, 1982: 415). A detailed study of trends by classes is underway, and preliminary indications are that the four classes require no major modifications of present conclusions.

Similarly, Woodman and Wayne (1985) raise the possibility that misleading results may be generated by including studies from the 1960s as well as the 1980s. Methodological sophistication and practical know-how presumably grow, and this may influence associations between rigor and outcomes. However, we know that QWL success rates vary only marginally over the 22-year period of observation (Golembiewski & Sun, 1989). In addition, preliminary analysis contrasting early QWL studies with later ones isolates no regular differences from present patterns.

References

Bass, B.M. 1983. Issues involved in relations between methodological rigor and reported outcomes in evaluations of organization development. *Journal of Applied Psychology,* 68: 197-199.

Bullock, R.J., & Svyantek, D.J. 1985. Analyzing meta-analysis. *Journal of Applied Psychology,* 70: 108-115.

Bullock, R.J., & Svyantek, D.J. 1983. Positive-findings bias in positive-findings bias research. In K.H. Chung (Ed.), *Proceedings of the Annual Meeting of the Academy of Management:* 221-224. Dallas, TX: Academy of Management.

Campbell, D.T., & Stanley, J.C. 1963. *Experimental and quasi-experimental designs for research.* Boston: Houghton-Mifflin.

Golembiewski, R.T., Proehl, C.W., Jr., & Sink, D. 1981. Success of OD applications in the public sector. *Public Administration Review,* 41: 679-682.

Golembiewski, R.T., Proehl, C.W., Jr., & Sink, D. 1982. Estimating the success of OD applications. *Training and Development Journal,* 72: 86-95.

Golembiewski, R.T., & Sun, B.C. 1988. QWL, one more time. *Healthcare Human Resource Forum,* 1: 1-2.

Golembiewski, R.T., & Sun, B.C. 1989. QWL improves worksite quality. *Human Resource Development Quarterly,* 1: 35-44.

Margulies, N., Wright, P.L., & Scholl, R.W. 1977. Organization Development techniques. *Group & Organization Studies,* 2: 439-441.

Mitchell, E. 1981. OE produces results. *OE Communique,* 5: 92-93.

Morrison, P. 1978. Evaluation in OD. *Group & Organization Studies,* 3: 42-70.

Nicholas, J.M. 1982. The comparative impact of Organization Development interventions on hard criteria measures. *Academy of Management Review,* 7: 531-542.

Skelley, B.D. 1989. Workplace democracy and OD. *Public Administration Quarterly,* 13: 176-195.

Sun, B.C. 1988. *Quality of working life programs.* Unpublished doctoral dissertation, University of Georgia, Athens, GA.

Terpstra, D.E. 1981. Relationship between methodological rigor and reported outcomes in Organization Development research. *Journal of Applied Psychology,* 66: 541-543.

Terpstra, D.E. 1982. Evaluating selected Organization Development interventions. *Group & Organization Studies,* 7: 402-417.

Woodman, R.W., & Wayne, S.J. 1985. An investigation of positive-findings bias in evaluation of Organization Development interventions. *Academy of Management Journal,* 28: 889-913.