QWL APPLICATIONS IN PUBLIC AGENCIES: DO SUCCESS RATES REFLECT A POSITIVE-FINDINGS BIAS?[*]

Robert T. Golembiewski
Department of Political Science
University of Georgia
Athens, Georgia 30602

Ben-chu Sun
Graduate Program of Public Administration
National Chengchi University
Mucha, Taipei
Taiwan, R.O.C.

Abstract

This article builds on reports of substantial success rates in Quality of Working Life (QWL) applications, and tests the criticism that these attractive outcomes merely reflect wimpish methodology. As the rigor of research designs increases, critical observers predict, so also will outcomes deteriorate. This panel of QWL applications, N = 44, does not support a positive-findings bias in public-sector applications, where QWL success rates approximate those in business applications.

Quality of Working Life (QWL) proponents make strong cases that management bottom-line needs can be met, while empowering employees and approaching representative political ideals in organizations, in both business and public sectors. For example, two

QWL theorists (1) propose that four dimensions
underlaying the humanization of work can and should
be attended to:

o   Security: the worker's freedom from fear
     and anxiety concerning his health, safety,
     income and future employment.

o   Equity: the worker receives compensation
     commensurate to his contribution to the
     value of a service or product.

o   Individualism: work should stimulate the
     development of the individual's unique
     abilities, craftsmanship, and the capacity
     for continued learning rather than force
     him into a mechanized role.

o   Democracy: . . . the worker [is not] a
     passive object . . . Wherever feasible,
     workers should manage themselves;
     authoritarian, hierarchical control should
     be replaced by cooperative self-managed
     groups.

Especially in the public-sector, however, some
observers dismiss such claims for both employees and
employing organization. Some see the claims as
illusions, and often cruel ones. Such observers
propose that public-sector features are refractory to
QWL values and designs, and that success rates are
low. Brief attention below will be devoted to QWL as
illusion, since other work provides details (2).

Basic attention here goes to QWL as artifactual,
with specific reference to the criticism that QWL
evaluative studies lack methodological and design
rigor. Consequently, despite substantial success
rates reported in over three dozen published surveys

of applications (3), some critics urge that a positive-findings bias operates to induce artifactually-high estimates of outcomes. As the rigor of evaluative studies increases, this criticism proposes, so also do outcomes become less attractive and positive. At times, this alleged bias is seen as buttressed by powerful reinforcers--e.g., that editors and their boards prefer to publish positive results.

This paper is part of a series testing a large panel of QWL applications for illusion and artifactuality (4). The panel is the largest presently available (N = 231), and details about its assemblage and character are available elsewhere (5).

FEATURES OF PUBLIC-SECTOR APPLICATION

The QWL sub-sample of interest here contains 44 cases, and its major features need early introduction. No completely-satisfactory way exists, short of reading each case in the panel of 231, to detail the characteristics of the 44 public-sector QWL applications. But Exhibit 1 presents substantial detail and four points highlight useful generalizations about the summary.

First, about two-thirds of the applications occur at sub-federal levels of government. In short, the federal government does not take the lead in disseminating QWL.

Second, QWL applications have been made in the major government areas or functions. Civil applications clearly dominate, but the panel includes a range of others--military, schools, and so on.

EXHIBIT 1

Features of Public-Sector QWL Applications

I.	**Locus**		**N =**
	1.	Federal	10
	2.	Sub-Federal, usually Local Government	32
	3.	Combined	2
II.	**Function Involved**		**N =**
	1.	Civil	33
	2.	Military	4
	3.	Police	1
	4.	Schools	2
	5.	Combined	4
III.	**Primary Level of Application**		**N =**
	1.	Operators	14
	2.	First-line supervisors and subordinates	5
	3.	Supervisors only	1
	4	Executives and mid-managers	2
	5.	Multiple	21
IV.	**Primary Designs**		**N =**
	1.	Alternative work schedules	9
	2.	Job enrichment	7
	3.	Feedback, usually using surveys	5
	4.	Union/management collaboration	2
	5.	Team building	3
	6.	Incentive plans	2
	7.	Management by objectives	2
	8.	Combined	12

Third, QWL applications tend to emphasize operating and supervisory levels, but Exhibit 1 shows that about half of the cases encompass multiple levels of organization. Specifically, this refers to the inclusion of three or more levels, typically including operators and their first-line supervisors.

Fourth, the 44 public-sector QWL applications include a broad representation of designs. About 30 percent of the applications get classified as Combined, which implies the useful intent of reinforcing and supplementing effects--as when

changes in policies, procedures, or structures are
employed to reinforce interaction-centered efforts
that highlight worksite inadequacies.

A brief review of the cases in which single
designs dominate serves two purposes. It not only
provides information about frequencies, but that
review also highlights the three major QWL approaches
to change--via interaction, structure, and policies
or procedures.

Of the single designs, alternative work
schedules--especially flexible workhours--are most
often relied on. As is widely known (6), such
programs generate attractive outcomes for both
employees and their organizations in a very high
percentage of cases. This intervention seeks to
increase responsible employee freedom via a change in
policy. Job enrichment is the second most popular
design, and it seeks the same end via structural
arrangements that permit the employee greater scope--
in skills, self-discipline, pacing, and so on.
Various feedback designs rank third in usage among
the single designs. They emphasize interaction that
generates valid and reliable information that permits
greater responsible freedom at work.

Broadly, this review of some major features of
44 QWL applications in the public sector implies that
the present panel has quite a comprehensive reach
that extends to different loci, includes major
public-sector areas or functions, encompasses
multiple levels, and involves the full range of QWL
interventions. This QWL reach implies that its
grasp--as reflected in substantial success rates--is
not an artifact of selectivity or narrow biases.

TABLE 1
QWL Success Rates on Hard-Criteria Outcomes,
Public Versus Business Sectors

QWL Outcomes, in percent

	N =	1. Highly Positive and Intended Effects	2. Definite Balance of Positive and Intended Effects	3. No Appreciable Effects	4. Negative Effects
Private Sector	185	71.4	20.5	4.9	3.2
Public Sector	44	59.1	27.3	9.1	4.6

PUBLIC-SECTOR QWL AS ILLUSION

Earlier papers (7), focus on success rates--that is, on whether QWL achieves its intended effects. Table 1 summarizes the data from a comparison of the public and private sectors. Note two points. Thus, success rate here is defined in terms of hard-criteria‚ outcomes--changes in objectively-assessed productivity, cost data, turnover, and so on. A similar table for "soft" measures--self-reported changes in attitudes, and the like--provides essentially the same results. Moreover, two cases could not be assigned to either of the two sectors. Hence N = 229.

Table 1 presents no great interpretative challenge and, on its face, supports two conclusions. Public-sector outcomes of QWL applications generate

substantial success rates. In sum, over 85 percent
of the cases are rated as having at least a definite
balance of positive and intended effects.   In
addition, the outcomes in the two sectors differ a
bit, but only a moderate advantage goes to business
applications.

Measurement details underlying Table 1 are
available elsewhere (8), but several summary
statements provide useful context for Table 1.
First, all codings in this and related reports derive
from three independent raters, all advanced doctoral
students in the behavioral sciences. Inter-observer
abilities vary a bit for different codings, but in
the range of the low .90s. So coding unreliability
presents no major problems.

Second, strict rules guide assessments of QWL
outcomes, with the typical evaluative study reporting
pre- versus post-scores on several variables. A "1"
assignment in Table 1 requires that changes in most
variables reported in a study are in the expected
direction--toward higher productivity and lower
costs--and are large in at least half the cases.
"Large" here means a change that achieves the .05
level or is greater than 10 percent. A "2"
assignment also requires that most changes are in the
expected direction, but fewer are large. Virtually
all cases containing large and unexpected changes get
a "4" code.

Third, in general, conservative scoring
conventions prevail. For example, if observers
disagree about an outcome code, discussion follows.
Failure to achieve agreement results in the
assignment of the lowest outcome score-in-contention.

Note that inter-observer reliabilities are calculated
before discussion.

PUBLIC-SECTOR QWL AS ARTIFACT

The outcomes in Table 1 are attractive, and even
formidable. But do they reflect less about
relationships in nature than they do about a
positive-findings bias (PFB)?

Testing for PFB rests on three methodological
emphases. Two measures of methodological rigor are
introduced; conventions about measuring the outcomes
of QWL applications are reviewed; and the properties
of statistical methods are sketched.

1. Two Measures of Rigor.

The summary of QWL outcomes in Table 1 tells us
nothing about the methodological quality of the 231
evaluative studies, and so the apparent success rates
there may reflect the inadequacies of research
design.

This possibility is taken very seriously here,
and three estimates are made of the rigor of each
evaluative study. Two separate estimates deal with
the rigor of methodology and design (M/D), while the
third estimate deals with the degree to which each
study successfully deals with threats to internal and
external validity (IEV). The correlation of the two
M/D measures is high--r=.91. So the assessments of
rigor admitted to analysis here include one M/D and
one IEV measure.

Woodman and Wayne's M/D Score

Based on criticisms (9) of Terpstra's (10)
original formulation, Woodman and Wayne (11) propose

a 0-9 measure of M/D rigor. Their nine dimensions
have been slightly modified by Sun (12), and assign 0
or 1 scores for:

1.   Sampling strategy: 1 indicates a full
     census or a representative sampling
     strategy like random sampling, stratified
     sampling, or cluster sampling. A
     nonrepresentative sampling strategy or an
     unspecified sampling plan is marked 0.

2.   Sample size: 1 is assigned when $N > 30$, 0
     when $N \leq 30$.

3.   Control or comparison groups: 1 indicates
     the presence of such a group, and 0
     indicates its absence.

4.   Random assignment utilization: 1
     represents use of random assignment, and 0
     its absence.

5.   Measurement strategy: 1 indicates
     longitudinal measurement, or measures taken
     at two or more times, and 0 indicates a
     cross-sectional or one-time measurement.

6.   Reliability and validity of measures
     employed: 1 represents reliability of
     measures $\geq .6$, along with some evidence of
     validity; 0 stands for no reliabilities
     reported, or reliabilities $< .6$ with no
     validity evidence.

7.   Criteria for dependent variable(s): 1
     reflects the use of some objective
     criteria; 0 stands for perceptual data
     only.

8.   Significance level: 1 indicates a
      probability of type I error $\leq$ .05; 0
      indicates p > .05.

9.   Use of a statistical analysis procedure: 1
      is yes, and 0 is no.

Morrison's I/E Score

Following Morrison (13), an estimate of internal
and external validity provides additional perspective
on the rigor of QWL studies. Her approach rests on
Campbell and Stanley (14), who identify four threats
to external validity and eight threats to internal
validity. In their usage, "external validity" refers
to the generalizability of results to other settings,
and "internal validity" refers to the likelihood that
the relationship between an independent and dependent
variable actually exists.

Morrison's I/E score varies from 0 to 12,
depending on whether or not each of the threats
described below is reduced. Details are available
elsewhere (15), but note here that the eight threats
to internal validity involve eliminating or reducing
these possible effects: non-treatment factors; aging
or development; sensitivity to testing instruments;
measurement instability in the instruments;
regression to the true mean; non-random selection;
experimental mortality over the period of
observation; and interaction of selection and
maturation.   In addition, Morrison's I/E score
assesses the design's management of four threats to
external validity: any reactive or interacting
effects of testing; interaction of selection effects
and the experimental variables; generalizability to

other settings; and multi-treatment interference.

A research design that reduces or eliminates each of the threats to validity is scored a 12. The poorest design is rated a 0.

2. Alternative Measures of Outcomes.

Each evaluative study is rated for two assessments of outcomes: a hard-criteria assessment; and a global assessment, which takes into account both "hard" and "soft" data. Moreover, due to the small number of applications with outcomes rated "3" and "4"--that is, as having no appreciable effect or negative effects--tests of rigor and outcome associations are run for both three- and four-category measures of outcomes. The three-category measure aggregates "3" and "4" ratings.

3. Statistical Methods.

All associations of rigor and outcomes will be tested by one-way analysis of variance, basically. In addition, any cases are shown by ANOVA to include non-random variance ($P \leq .05$) will be subjected to an analysis of all possible paired-comparison to assess both the magnitude as well as the direction of all differences. The Least Significant Difference test, as modified for unequal sub-sample sizes, is used for this purpose.

In sum, perfect support for the positive-findings bias requires rather stringent conditions. Thus each rigor and outcome association would have to show statistically-significant variance by ANOVA. Moreover, every possible pair of rigor and outcome combinations would be in a direction consistent with

TABLE 2
Illustrative ANOVA, QWL Outcomes and I/E Validity,
Hard-Criteria Assessment, Public-Sector Applications

	I/E Validity Scores			
	N	Mean	Standard Deviation	ANOVA
<u>Outcomes</u>				
1. Highly positive and intended effects	26	6.0	3.52	
				F=3.21
2. Definite balance of positive and intended effects	12	4.7	2.84	
				F= .033
3. No appreciable effects	4	9.8	2.50	
4. Negative effects	2	2.5	0.71	

the positive-findings bias. Specifically, this means
that as rigor increases, the outcomes deteriorate.
Finally, <u>perfect</u> support of the positive-findings
bias requires that each paired-comparison is
statistically significant as well as in the
appropriate direction.

Testing for Positive-Findings Bias

This analytic procedure is not complicated, as
Table 2 illustrates for one of the 8 tests of the PFB
hypothesis. There, ANOVA indicates that non-random
variation does exist in the association between
Morrison's measure of internal/external validity and
QWL outcomes. In addition, a substantial proportion

of the variance in scores is explained (eta^2 = .19).
However, the direction of the associations is not
consistent with the positive-findings bias. In fact,
only two of the six possible paired-comparisons fall
in the appropriate direction--(1) versus (3), and (2)
versus (3). Only in one-third of the cases, then,
are QWL outcomes less attractive when rigor is
greater. Moreover, although both paired-comparisons
directionally consistent with the PBF hypothesis do
attain $P \leq .05$, one of the four contrary cases also
attains statistical significance.

In sum, the case illustrated in Table 2 provides
middling and mixed support for the positive-findings
bias hypothesis.

The eight total tests of the PFB hypothesis--
encompassing two measure of rigor, hard-criteria as
well as global assessments of outcomes, and three- as
well as four-category estimates of outcomes--provide
even less support for a positive-findings bias in
public-sector QWL applications. See Table 3 for
summary details. Paramountly, only the ANOVA run
illustrated in Table 2 achieves statistical
significance, so seven of eight distributions of
outcome and rigor are random only. Note also that
the directions of paired-comparisons--the most gentle
test of the positive-findings bias--are in relatively
equal proportions consistent with the hypothesis, and
contrary to it.

DISCUSSION

In sum, this paper shows that a positive-
findings bias contributes very little to the definite
balance of attractive QWL outcomes in the public-

TABLE 3
Summary Table, Tests of Positive-Findings Bias, Public Sector Applications

		Raw Data	Percent
1.	Overall ANOVAs attain P ≤ .05	1 of 8 cases	12.5
2.	Paired-comparisons in direction consistent with hypothesis and statistically significant by LSD test, Modified	2 of 36 cases	5.6
3.	Paired-comparisons in direction consistent with hypothesis	21 of 36 cases	58.3
4.	Paired-comparisons in direction contrary to hypothesis	15 of 36 cases	41.7
5.	Paired-comparisons in direction contrary to hypothesis and statistically significant by LSD test, Modified	1 of 36 cases	2.8

sector. QWL outcomes do <u>not</u> deteriorate with
heightening rigor.

This finding encourages QWL applications in the
public sector, and also gets reinforcement from two
other sources. Thus, public-sector applications seem
to have virtually the same pattern of success rates
as do business applications. See Table 1. Moreover,
the present public-sector panel encompasses four
kinds of QWL interventions that are widely
applicable. These four types of QWL interventions
include:

o   Human Processual, which seeks to improve
         personal and group functioning, as by team-
         building

o   Sociotechnical Systems, which design micro-
        organizations to integrate technological
        and social systems to create flexible
        structures and high personal involvement,
        in autonomous teams.
o   Technostructural, which can apply at macro-
        as well as micro-levels--e.g., workflow
        reorganization and matrix models as well as
        job enrichment.
o   . Combined, which rely on interventions from
        two or more of the above classes.

Real tethers exist on present findings, however.
Thus, the population of public-sector interventions
is not large (N = 44), although it represents the
largest available panel. Subsequent studies of
larger populations may not replicate the present
pattern of results. In addition, studies in the
present panel typically are pre-test versus short
post-test designs, and longitudinal studies may find
that effects decay quickly.

This research also cannot be definitive about
two other major issues, but these do nothing to
discourage QWL interventions. First, the basic
hypothesis here may be misguided. That is, the
positive-findings bias assumes that rigorous research
designs will tend to isolate outcomes rated (3) and
(4)--no appreciable effects and negative effects,
respectively. In contrast, it may be that negative
effects will be quite obvious to all but the most
self-interested observers. If so, research rigor is
arguably not crucial, except perhaps to differentiate
outcomes (1) from (2).

Second, a kind of dual and simultaneous incompetence may exist. An incompetent plan for application often may be paired with an inelegant design for research. This possibility sets aside the positive-findings bias. Here, the problem is not in QWL, but is relocated to the intervenor/researcher.

NOTES

*

An earlier version of this paper was presented at the Annual Meeting, American Political Science Association, September 3, 1989.

REFERENCES

(1)   Herrick, Neal, Q., and Maccoby, Michael. "Humanizing Work," in Davis, L.E. and Cherns, A.B., (editors), The Quality of Working Life, The Free Press, New York, 1975, pp. 63-77.

(2)   Sun, Ben-chu. Quality of Working Life Programs. Unpublished doctoral dissertation, University of Georgia, Athens, Georgia, 1988; Golembiewski, Robert T., and Sun, Ben-chu. "Enriching Work and Empowering Employees." Paper presented at Annual Meeting, American Society for Public Administration, Miami, Florida, April 1989; Golembiewski, Robert T., and Sun, Ben-chu. "QWL Improves Worksite Quality." Human Resources Development Quarterly 1 (September 1989): in press; Golembiewski, Robert T., and Sun, Ben-chu. "Positive-findings bias in QWL research: A Comparison of Public and Business Sectors." Public Productivity Review (1990): in press.

(3)   Sun, Ben-chu. Quality of Working Life Programs, 1988.

(4)   Golembiewski, Robert T., and Sun, Ben-chu. "Enriching Work and Empowering Employees," 1989; Golembiewski, Robert T., and Sun, Ben-chu. "QWL Improves Worksite Quality," September 1989.

(5) Sun, Ben-chu. Quality of Working Life Programs,
1988.

(6) Golembiewski, Robert T., and Proehl, Carl W.,
Jr. "Public-Sector Applications of Flexible
Workhours." Public Administration Review 40: 72-85.

(7) Golembiewski, Robert T., and Sun, Ben-chu.
"Enriching Work and Empowering Employees," April
1989; Golembiewski, Robert T., and Sun, Ben-chu.
"QWL Improves Worksite Quality," September 1989.

(8) Sun, Ben-chu. Quality of Working Life Programs,
1988.

(9) Bullock, R.J., and Svyantek, D.J. "Positive-
Findings Bias in Positive-Findings Bias Research."
In Chung, K.H., (editor), Proceedings: 221-224.
Annual Meeting of the Academy of Management, Dallas,
Texas, 1983.

(10) Terpstra, David E. "Relationship Between
Methodological Rigor and Reported Outcomes in
Organization Development Evaluation Research."
Journal of Applied Psychology 66 (December 1981):
541-543,

(11) Woodman, Richard W., and Wayne, Sandy J. "An
Investigation of Positive-Findings Bias in Evaluation
of Organization Development Interventions." Academy
of Management Journal 28 (December 1985): 889-913.

(12) Sun, Ben-chu. Quality of Working Life Programs,
1988.

(13) Morrison, Peggy. "Evaluation in OD." Group &
Organization Studies 3 (March 1978): 42-70.

(14) Campbell, Donald, T., and Stanley, J.C.
Experimental and Quasi-Experimental Designs for
Research. Houghton Mifflin, Boston, Massachusetts,
1963.

(15) Sun, Ben-chu. Quality of Working Life Programs,
1988.