Bulletin of the Psychonomic Society
1983, Vol. 21(1), 69-72

Paradox lost, paradox regained: Reply from a flagellated straw man

JAMES V. BRADLEY
New Mexico State University, Las Cruces, New Mexico 88003

The optimal-pessimal paradox (Bradley, 1975) has been criticized on bizarre grounds by Childs (1980). Assumptions that it never made were attributed to it and attacked. Empirical evidence for its existence (which occupied a large portion of the criticized article) was totally ignored, and it was treated as a mere theoretically based conjecture. Originally proposed solutions to the problems it presents were dismissed and replaced by ineffectual alternatives. In spite of Childs' claim that "the paradox, although theoretically sound, is grounded upon assumptions that are empirically untenable," the paradox makes no such assumptions and is an empirical fact.

The optimal-pessimal paradox (Bradley, 1975) and the hazards it presents to psychological research have been disputed by Childs (1980).[1] His argument appears to rest upon three major points, all of which collapse under scrutiny: (1) His principal basis for disputing the paradox appears to be the contention that it is based upon certain alleged assumptions that are empirically untrue or rarely met. Actually, the paradox makes no such assumptions. Childs creatively obtained these "assumptions" by selecting certain highly specific input conditions for a fictitious illustrative example in the criticized article, ignoring others (which did not lend themselves as plausibly to his attack), restating the selected conditions in highly generalized terms, and then alleging that these generalizations were "assumptions" of the paradox, although no such assumptions had been stated or implied for the paradox either by Bradley or by the logic of the situation. (2) He treats the paradox as though it were merely a speculative theory unsupported by empirical evidence. In doing so (although stressing actual research conditions throughout his article), he totally ignores the strong empirical evidence given in the criticized article both for the existence of the paradox in actual psychological experimentation and for the extreme nonrobustness resulting from it. Rather than being a theory unsupported by empirical observations, the paradox is an empirically discovered and established fact that is supported by theoretical considerations (Bradley, 1975, 1977, 1982) that sanction its generalization. (3) Finally, he implies that in those situations in which the hazards (of L-shaped populations and consequent nonrobustness) do occur, there are more and better solutions than those outlined in the criticized article. Yet the alternative solutions proposed by Childs are largely ineffectual, and some will not work at all. These points will be elaborated upon below. (There are many fallacies in Childs' article; only the major ones will be answered.)

THE PARADOX AND THE EVIDENCE FOR IT

Consider a timed task subject to errors, each of which causes an increment in total task time. As the task becomes easier (or as its performance becomes more skilled), errors become less probable so that more and more of the time-score distribution becomes concentrated over the low time scores associated with zero errors and less and less of it lingers over the higher time scores associated with the commission of various numbers of errors. Consequently, the distribution, which may have been quasi-normal when errors were frequent, eventually becomes roughly L-shaped as error frequency diminishes. So, as the task becomes easier (or as its performance improves), the shape of the time-score distribution violates the normality assumption of parametric statistics more and more seriously. That is, as experimental task conditions become more nearly optimal, parametric statistical testing conditions tend to become more nearly pessimal. This is the optimal-pessimal paradox.

Although the increasing L-shapedness is virtually a logical consequence of the stated conditions, the paradox was not discovered through theoretical considerations. Rather, it was first encountered as an empirical fact. The increasing skewness of error-increasing-time-score distributions with increasingly favorable experimental conditions (see Figures 2-5 in Study 1 of Bradley, 1968b, or Figure C of Bradley, 1977, or Figure 1 of Bradley, 1982) was discovered empirically in data taken under conditions characterizing a routine experiment in engineering psychology (Bradley & Wallis, 1959) and quickly confirmed in the experiments of others (see Figure D of Bradley, 1977). Empirical sampling studies, using either these very same empirical L-shaped distributions of time scores or highly similar L-shaped distributions as the sampled population, then showed alarming degrees of nonrobustness associated with L-shaped

populations (Bradley, 1968a, 1968b Study 6, 1971,[2] 1975, 1976, 1978, 1980a, 1980b, 1980c; Wike & Church, 1982). Further empirical sampling studies showed increasing nonrobustness with increasing degrees of skewness or L-shapedness (Bradley, 1968b Study 7, 1973,[2] 1976). Some of these empirical data were conspicuously presented (e.g., in Figure 4 and Tables 1 and 2, Bradley, 1975) in the criticized article, but they were ignored by Childs (1980), who seems to regard the paradox as merely a theoretically based conjecture.

Statistical and mathematical considerations (Bradley, 1975, 1977, 1982) supplement these empirical facts by showing that the specific empirical results can be validly generalized to a much broader range of conditions than those under which they were discovered. It is a well known fact (for which some references were given in the criticized article) that the robustness of parametric tests on means tends to diminish rapidly as population skewness and/or kurtosis increases. Bradley (1982) presents mathematically derived curves showing that the coefficients of skewness and kurtosis of an L-shaped distribution increase dramatically as the long positive tail becomes thinner and thinner. These curves therefore support the conclusion that as the task becomes easier than it was in the already investigated situations for which empirical data are available, parametric tests should become much more nonrobust than the spectacular degree already found (when the coefficients of skewness and kurtosis never exceeded 3.42 and 17.29, respectively). Thus, the optimal-pessimal paradox was discovered and established empirically in specific instances, whereas its generalizability is warranted by various theoretical considerations.

ALLEGED ASSUMPTIONS OF THE PARADOX

As apparently his main argument, Childs (1980) alleges that "the paradox itself is based upon at least four assumptions not applicable to many applied research settings" (p. 117) and, indeed, that are "empirically untenable" (p. 113). Childs makes this claim because he has failed to distinguish between the optimal-pessimal paradox and the particular features of an example illustrating it. He has fallaciously identified the conditions qualifying and accompanying the example as "assumptions" of the paradox. (He also attacked some of these conditions as "unrealistic" while at the same time acknowledging that they apply to a "hypothetical" illustration.) The manner in which he arrived at this fallacious conclusion can be seen as follows.

In order to show exactly how increasingly rare errors can lead to increasingly skewed time-score distributions, the criticized article resorted to a fictitious example, introduced by the words "The influence of performance conditions upon the shape of the time-score distribution can best be *illustrated* by 'building' such a distribution in several successive stages, i.e., by a sort of logical synthesis" (Bradley, 1975, p. 322; italics added). This "building" task necessitated the elaborate specification of input conditions (which determined the probabilities and time scores plotted in the first three figures of the criticized article). In specifying them, an attempt was made to choose the least complicated and most easily conceptualized set of conditions that would produce parameter relationships roughly in harmony with those that had actually been obtained for the skewed empirical distribution of time scores shown in the last figure of the criticized article. (For example, since the errors accompanying the time scores were well fitted by a Poisson distribution, input conditions were chosen under which errors would have a Poisson distribution; see "a," below.)

Thus it was "supposed that" for the 100-yard dash, (a) "a fall is equally likely to occur at any point and that there is no limit to the number of possible falls," (b) "each error committed consumes the same amount of time . . . on the average, and therefore increases running time by equal increments," (c) "the time to run the 100-yard . . . dash is the sum of the time to run the 100 yards . . . without falling and the times consumed in each fall, and that these component times are all independent of each other and normally distributed," (d) "the standard deviation of time scores for errorless trials is $\sigma_0 = \overline{\Delta t}/5$ and that the standard deviation of the time consumed in falling is $\sigma_f = \overline{\Delta t}/2$, where $\overline{\Delta t}$ is the constant distance between adjacent $\overline{t_i}$'s caused by the constant average amount of time consumed by each fall" (Bradley, 1975, pp. 322-324). The example also let the average number of errors be .1 for a "skilled runner" and 5 for an "inept runner." Notice that all these conditions are stated in particularistic terms, specifically with references to running a 100-yard dash. Nowhere in the article was it stated or implied that these or similar conditions were "assumptions" of the optimal-pessimal paradox or anything other than specific input conditions for an easily conceptualized, fictitious, illustrative example showing "how and why a more or less bell-shaped time-score distribution tends to be associated with frequent time-increasing errors and an L-shaped distribution with rare errors" (Bradley, 1975, p. 324).

It is astonishing, therefore, to read in Childs' (1980, p. 113) article that "the optimal-pessimal paradox includes the following assumptions: (1) Error probabilities are equal across all task segments, and therefore may be fitted to the Poisson distribution (Bradley, 1975, p. 322). (2) Error commissions uniformly increase task execution times (p. 322). (3) Component error times are orthogonal (p. 323). (4) Robustness of parametric tests is greatly reduced by skewness (p. 326)."

Childs' (1980) "Assumptions" 1, 2, and 3 are totally unwarranted generalizations of the specific input Conditions a, b, and c of the fictitious illustrative example to the entire optimal-pessimal paradox. Not only has he replaced the specific language about "falls," and so on,

with general terms such as "task segments," but he has entirely omitted conditions that (to be consistent) should surely qualify as part of the unjustifiably generalized "assumptions." What happened to "Assumption" d? And how did the specification in "Assumption" c that each of the component times is "normally distributed" drop completely out of Childs' translation of the original "Assumption" c into his "Assumption" 3? These curious omissions, if included, would have vitiated his point.

Childs (1980) lists as the paradox's fourth alleged "assumption" that "robustness of parametric tests is greatly reduced by skewness," although it is a known fact in support of which the criticized article gave both references and empirical robustness data (in Table 2 of Bradley, 1975) based upon samples drawn from the very same empirical population (Figure 4 of Bradley, 1975) whose existence and skewness Childs persists in ignoring. He tacitly rejects all this, preferring to cite in refutation Lindquist's (1953) rather biased (see Bradley, 1978) account of Norton's 1952 study and a 1968 secondary source (Kirk, 1968) largely echoing Lindquist. Norton's study investigated only the F test and did so under mild conditions generally favorable to robustness. His populations were artificial (mathematical density functions), were not L-shaped (none showed an abrupt ascent to the mode followed by a sharply precipitous descent to a very long and shallow positive tail), and were much less skewed than the L-shaped populations obtained so often under conditions that give rise to the paradox. Norton's results are therefore irrelevant to the paradox. As mentioned earlier, robustness studies using such L-shaped populations (Bradley, 1968a, 1968b, 1971, 1973, 1975, 1976, 1978, 1980a, 1980b, 1980c; Wike & Church, 1982) show devastating nonrobustness, and they do so under a variety of not easily summarized conditions.

FINESSING THE PROBLEM

In discussing ways of avoiding the dangers attendant upon the paradox, the criticized article (Bradley, 1975) suggested various ways of circumventing the problem, the most important of which was to use distribution-free statistics. (The efficacy of this solution has recently been shown by Blair and Higgins, 1981.) Childs (1980) has a different set of suggestions, which are largely ineffectual: employ multivariate techniques, use transformations, or use repeated-measures designs. Multivariate techniques that made no assumptions about the shape of the distribution would, of course, finesse the problem (by being distribution free), but not necessarily otherwise. Increasing the power of an experiment by using repeated-measures designs does not have any necessary influence upon robustness of Type I error probability against nonnormality, and this type of robustness is just as desirable as power robustness. Furthermore, the power of parametric tests under nonnormality may be far inferior both to their own power under normality and to the power of a nonparametric competitor under the same nonnormal condition (Blair, Higgins, & Smitley, 1980). It would seem, then, that increasing the power of parametric tests by using repeated measures would finesse the problem only if one knew in advance that the null hypothesis was false (in which case, there would be no need to do the experiment) and that the test would have a power of nearly 1.00 (in which case, graphs of the data would probably be sufficiently convincing without any statistical test).

Transformations applied to L-shaped populations simply cannot alter the fact that the mode is at the extreme end of the distribution, and this fact hinders the transformations from greatly reducing skewness. The effectiveness of the same three transformations mentioned by Childs (1980) as "useful for normalizing positively skewed distributions" [$X' = \log(X + 1)$ and $X' = \sqrt{X + (1/2)}$] and "for normalizing response time distributions" ($X' = 1/X$) was thoroughly checked out by Bradley (1982). When applied to L-shaped distributions (including the empirical L-shaped distribution prominently displayed in the criticized article but ignored by Childs), the recommended transformations "(1) were highly unsuccessful in most cases and never were completely satisfactory, (2) were least successful in those cases in which they were most needed (i.e., when L-shapedness was greatest), and (3) when most successful, were most adversely influenced by the location of the distribution, which in turn may be a function of strictly arbitrary or fortuitous experimental conditions (such as . . . the distance through which the subject's hand must travel to operate the push button). It would seem, then, that transformations can hardly be relied upon to solve the problem of L-shaped distributions" (Bradley, 1982, p. 86). Wike and Church (1982) have also applied log, square-root, and reciprocal transformations to one of Bradley's L-shaped populations and found them ineffectual. Transformations work much better upon the mildly skewed distributions having the lopsided hump shapes so familiarly displayed in elementary statistical textbooks. But the more skewed distributions accompanying the optimal-pessimal paradox tend to be L-shaped.

THE PROBLEM TRANSCENDS THE PARADOX

The paradox is restricted to a particular type of experimental situation in which the dependent variable is time scores that are increased by the occurrence of errors. The problem posed by it is that when errors are infrequent, the distribution of scores tends to become L-shaped and that such highly skewed distributions are extremely conducive to nonrobustness. However, badly skewed L-shaped distributions can occur for a large variety of reasons under many diverse situations (Blair, 1979; Bradley, 1977), only one of which is the situation to which the paradox refers. Therefore, the problem created by the hazard of L-shaped populations and con-

sequent nonrobustness is considerably more likely to be encountered than is the optimal-pessimal paradox, which is simply a subcategory of the larger problem.

SUMMARY

The optimal-pessimal paradox applies when the dependent variable is time scores for a task subject to errors that delay its completion. The paradox is not further encumbered by fragile and unlikely conditions of the type imagined by Childs (1980). Furthermore, L-shaped distributions that may have devastating effects upon robustness can occur under a variety of experimental circumstances, only one of which is that involved in the paradox. So the hazards associated with the paradox actually transcend the paradox itself. The optimal-pessimal paradox is inconvenient, vexatious, and embarrassing, but it is an empirical fact that cannot be dismissed by treating it as a conjecture or attacking fictitious "assumptions."

REFERENCES

BLAIR, R. C. Comment on the relative usefulness of the two independent means t-test and Wilcoxon's rank-sum test for the analysis of educational data. *Florida Journal of Educational Research,* 1979, **21**, 97-110.

BLAIR, R. C., & HIGGINS, J. J. A note on the asymptotic relative efficiency of the Wilcoxon rank-sum test relative to the independent means t test under mixtures of two normal distributions. *British Journal of Mathematical and Statistical Psychology,* 1981, **34**, 124-128.

BLAIR, R. C., HIGGINS, J. J., & SMITLEY, W. D. S. On the relative power of the U and t tests. *British Journal of Mathematical and Statistical Psychology,* 1980, **33**, 114-120.

BRADLEY, J. V. *Distribution-free statistical tests.* Englewood Cliffs, N.J: Prentice-Hall, 1968. (a)

BRADLEY, J. V. Studies in research methodology. *Dissertation Abstracts* (Monograph Section), 1968, **28**, 4815B-4816B. (b)

BRADLEY, J. V. A large-scale sampling study of the central limit effect. *Journal of Quality Technology,* 1971, **3**, 51-68.

BRADLEY, J. V. The central limit effect for a variety of populations and the influence of population moments. *Journal of Quality Technology,* 1973, **5**, 171-177.

BRADLEY, J. V. The optimal-pessimal paradox. *Human Factors,* 1975, **17**, 321-327.

BRADLEY, J. V. *Probability; decision; statistics.* Englewood Cliffs, N.J: Prentice-Hall, 1976.

BRADLEY, J. V. A common situation conducive to bizarre distribution shapes. *American Statistician,* 1977, **31**, 147-150.

BRADLEY, J. V. Robustness? *British Journal of Mathematical and Statistical Psychology,* 1978, **31**, 144-152.

BRADLEY, J. V. Nonrobustness in classical tests on means and variances: A large-scale sampling study. *Bulletin of the Psychonomic Society,* 1980, **15**, 275-278. (a)

BRADLEY, J. V. Nonrobustness in one-sample Z and t tests: A large-scale sampling study. *Bulletin of the Psychonomic Society,* 1980, **15**, 29-32. (b)

BRADLEY, J. V. Nonrobustness in Z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society,* 1980, **16**, 333-336. (c)

BRADLEY, J. V. The insidious L-shaped distribution. *Bulletin of the Psychonomic Society,* 1982, **20**, 85-88.

BRADLEY, J. V., & WALLIS, R. A. Spacing of push button on-off controls. *Engineering and Industrial Psychology,* 1959, **1**, 107-119.

CHILDS, J. M. Time and error measures of human performance: A note on Bradley's optimal-pessimal paradox. *Human Factors,* 1980, **22**, 113-117.

KIRK, R. E. *Experimental design: Procedures for the behavioral sciences.* Belmont, Calif: Brooks/Cole, 1968.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

NORTON, D. W. *An empirical investigation of the effects of nonnormality and heterogeneity upon the F-test of analysis of variance.* Unpublished doctoral thesis, State University of Iowa, 1952.

WIKE, E. L., & CHURCH, J. D. Nonrobustness in F tests: 1. A replication and extension of Bradley's study. *Bulletin of the Psychonomic Society,* 1982, **20**, 165-167.

NOTES

1. The most appropriate time and place for a reply are immediately and in the same journal in which one has been attacked. A reply to Childs was submitted to *Human Factors* in September 1980 and accepted, after revision. However, it was subsequently withdrawn from publication due to disputes over postacceptance editorial changes that I felt vitiated the reply.

2. These studies (Bradley, 1971, 1973) of the central limit effect upon the standardized sample mean, $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$, may be equally regarded as studies of the robustness of the one-sample Z test against nonnormality, since (under a given violation of the normality assumption) that robustness is entirely attributable to the central limit effect.

(Received for publication December 16, 1982.)