Published in *American Educational Research Journal* 1969, Vol 6 (1), pp. 112–116

Review of Lord and Novick

Lord, Frederic M., and Novick, Melvin R. (with contributions by Allan Birnbaum). *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison-Wesley, 1968. xii + 568 pp. $14.95.

After nearly a generation of drought, a torrent of test-theory texts has burst upon us: books by Ghiselli, Helmstadter, Magnusson, Horst, Nunnally, and myself all having appeared within the past few years. The advent of Lord and Novick into this distinguished company now brings a bracingly fresh outlook upon the future. For unlike previous text material in this area, most of which can be anthologized under the heading “Conventional Thoughts on Traditional Topics,” the present work well merits the subtitle “Research Frontiers in Assessment Theory.” In a few brief chapters, L&N deftly recapitulate the main theorems of classical test theory; from there it's off into the wilds of problems and concepts not previously domesticated between book covers. The result is a major contribution to advanced psychometrics—not because the authors report much that is entirely new but because they clarify and integrate important but heretofore disconnected recent developments. This work will undoubtedly soon become recognized as the definitive reference for students and specialists wishing to survey the advanced outposts of contemporary test theory.

Anyone contemplating use of the book as a primary teaching text will do well to heed the authors' own note of caution in this regard. L&N have managed to keep their mathematical exposition simple, terse, and remarkably lucid, but only at a price: the more difficult results are often stated without proof (though seldom without indicating where this can be found), and the reader is required to have a good working grasp of elementary mathematical statistics, sporadically augmented by some background in analysis of variance, calculus, and matrix algebra. And presentation of the more standard test-theoretic materials is often so condensed that the student who has not already worked through these ideas elsewhere will be hard pressed to know what to make of them here. There can be no quarrel with the authors' choice of expository tactics—by their fruits shall we judge them and

these fruits are most tasty. It only means that to give students a balanced diet this book should be supplemented by the peas and potatoes of a good traditional text. Of the three major developments reviewed by L&N, the most topical is the notion of “generic” reliability. Initially introduced by Cronbach and his collaborators a few years ago as the “theory of generalizability,” this concept envisions obtaining a subject's score on a generic test X by assessing him on some procedure X_k drawn randomly from a domain $\{X_k\}$ of specific test alternatives. Then the subject's true score on X—his "generic" true score—is a composite of his specific true scores on the various alternatives in $\{X_k\}$ and reliability theory for this test domain can explore the relations of the specific tests X_k to this generic true-score variable as well as to their own specific true-score components. Although previous contributions to this development have left its conceptual basis more than a little obscure, L&N's account together with the excellent recent article by Hunter *(Psychometrika,* 33: 1-18; 1968) now make its foundations pellucid.

Because generic reliability is still so invitingly new a frontier, so modest in its mathematical demands, and so seemingly practical in its promise to free reliability theory of its traditional but unrealistic parallel-forms presuppositions, it is a safe bet that this territory will soon be overrun by a herd of research prospectors. I predict, however, that when the dust has settled, little benefit will be found to have come of it all. For the relation of a generic test X to one of its specifics, X_k, is merely that of a test in whose definition a source S of error variance is uncontrolled to what the test becomes when S is standardized at some fixed value s_k. This relation is admittedly worth delving for theoretical insights, but to date the trend has been an emphasis on practicalities of parameter estimation at the expense of conceptual penetration. For applied testing, however, if we have (*a*) enough background data to estimate the reliability/validity parameters of test X with S held constant at s_k, while (*b*) subject *i*’s score on X is also known to have been obtained under condition s_k, then the classical theory of test X_k tells how best to interpret *i*’s score; whereas if either (a) or *(b)* is lacking, the classical theory of test X with S unstandardized again extracts all there is to learn from *i*’s score. In no way do I intend to derogate L&N's account of generic reliability, for they have done, and done well, only what well needed doing. But persons planning to leap aboard this particular bandwagon should be warned that they will

be lucky to reach the outskirts of town.

The second major thrust of contemporary test theory summarized by L&N with none of the fanfare it deserves is the senior author's own pioneer work on estimating the joint distribution of observed and true scores of a test when this distribution is not presumed to be bivariate normal. My own enthusiasm for this development lies in a poorly restrained impatience with traditional model-building attitudes which tolerate no end of implausible assumptions so long as these are mathematically seductive. Strong axioms are a perfectly respectable and perhaps necessary way to initiate command of a new problem area, but once the initial theoretic regimen has become well established attention should then turn to analytic and empirical probing of its assumptions to see how far these can be relaxed without essential loss and, where robustness is lacking, how severely the postulated ideals diverge from reality. Lord is essentially the first test theorist to have taken this second step toward maturity even if, to be sure, classical test theory has not drawn heavily on normality assumptions and his approach has some unpleasant presuppositional problems of its own.

Finally, L&N make amply clear the continuity between reliability theory and inferential factor analysis by filling the gap with an assortment of nonlinear factorial decomposition models of test data. Much of this “latent trait” material is contained in four chapters by Birnbaum which are noteworthy on two grounds: (1) they sharply point up, by contrast, the lucidity of L&N's own writing, and (2) they develop a provocatively novel approach to latent-trait estimation based on some of the more sophisticated concepts of inferential statistics. (The utility of this approach is debatable, however, for it ignores the increased efficiency of trait estimation afforded by information about the latent-trait variable's unconditional distribution. This is like estimating a statistical parameter by classical methods when known prior probabilities make a Bayesian argument feasible, or approximating a subject's true score on a test by his uncorrected observed score when knowledge of the test's reliability permits a regression estimate.) Unfortunately, all the specific latent-trait models presented here are one-factor idealizations whose prospects, if any, for generalization to a more plausible latent-trait space of empirically determined dimensionality are left undisclosed.

So far, my praise for this work has been only faintly stinted. But great virtues often cohabit with great defects, and this otherwise splendid achievement is disfigured throughout by one monumental sin of omission: a systematic refusal to think about the problem of correlated measurement errors. One of test theory's hoariest traditions is the “local independence” assumption that a person's distribution of measurement errors on one test unit is statistically independent of his error distribution on any other. Although this premise is absurdly unrealistic, especially when the measurements are taken in close temporal proximity to one another, L&N have relentlessly exploited local independence with never a hint that the results so obtained might thereby suffer from irrelevance to reality. That L&N's personal interests should run to models based on the local independence axiom is uncontrovertibly their privilege, but when these have practical implications for test design and score processing it is dubious wisdom to let such conclusions stand without cautionary qualms for whether they provide even a good first approximation to what should be said if measurement errors are appreciably correlated. As it is, the analytic consequences and empirical magnitude of correlated errors deserve emphasis as a fourth primary research frontier in test theory. Development of techniques to minimize correlated-error effects may prove to be one of the few broad-spectrum methods we have for enhancing test reliability (see Rozeboom, W. W., *Foundations of the Theory of Prediction*, pp. 441 ff.).

Addison-Wesley's production staff deserves special commendation for the visual appeal they have built into this volume. Its typeface gives pleasure to the eye and its strikingly handsome cover is far above the esthetic norm for academic books today. Even if you couldn't care less about test theory, buy a copy to give your office decor a touch of elegance.

William W. Rozeboom *University of Alberta*