It's Up for Debate: A Review of *Seminal Ideas and Controversies in Statistics* by Roderick J. A. Little

Gabriel Wallin

School of Mathematical Sciences, Lancaster University

Little, R. J. A. (2025). *Seminal ideas and controversies in statistics.* CRC Press, Boca Raton, FL. ISBN 9781032493565 (pbk).

In *Seminal Ideas and Controversies in Statistics*, Roderick J. A. Little examines influential papers and debates that have shaped modern statistics. Rather than providing a conventional history of statistics, the book organizes topics by foundational debates, computational advances, and design principles. Each chapter centers on a seminal article (or group of articles) that marked a controversy or turning point. Little provides historical context and technical exposition without excessive mathematical detail, and he reflects on the strengths, weaknesses, and unresolved questions surrounding each contribution.

The book is divided into three parts, which can be read in any order. Part I covers statistical inference foundations: Fisher's maximum likelihood formulation, the debate over tests of significance for 2×2 contingency tables, tensions between Fisher's significance testing and Neyman-Pearson hypothesis testing, fiducial inference through the Behrens-Fisher problem, and Birnbaum's likelihood principle with its controversies.

Part II examines methods that transformed practice: shrinkage estimation via Stein's paradox and empirical Bayes, alternatives to least squares including ridge regression and the lasso, multiple comparisons and false discovery rate control, generalized estimating equations

for longitudinal and clustered data, computational advances through the bootstrap and Bayesian MCMC, and the emergence of exploratory data analysis and data science.

Part III addresses design principles: randomization in survey sampling and design-based and model-based inference, randomized clinical trials through the Neyman/Rubin causal model, and propensity score methods for observational studies. Lastly, Little offers twenty writing tips for statistical communication in the Appendix.

This book lays out a range of perspectives and often indicates which the author finds most convincing. Little not only shares his own reflections but also draws on comments and debates from other leading figures in the field. For example, he compares advantages and drawbacks of frequentist and Bayesian approaches and is transparent about his own leanings. As he notes (p. 72), some criticisms of Bayesianism do not persuade him; at the same time, he emphasizes that *models are always wrong, and bad models lead to bad answers* (p. 73). A caution (p. 69) is that good repeated-sampling performance does not by itself guarantee that a particular inference is appropriate for the data actually observed. I believe this pragmatic stance aligns rather well with traditions in educational and behavioral statistics, where principles are often selected for their ability to yield sensible analyses in applied settings, and then validated against both theory and empirical evidence.

Many chapters are built around discussion papers, and Little uses those discussions to show how rather sharp exchanges can be productive. Following this model, journals often stage similar debates by inviting commentaries from scholars with contrasting views. The exchange between Gelman and discussants on Bayesian and frequentist reasoning is one example (Gelman, 2008; Wasserman, 2008). In the area of educational assessment, we have also seen a lively exchange on the theoretical foundations of observed-score equating, following van der Linden's call for a clearer test-theoretic basis (van der Linden, 2013; Holland, 2013). Little's book demonstrates why such debates matter. In that tradition, the book is not a one-way communication; Little concludes each chapter with discussion points that encourage readers to reflect on their own views of the issues raised.

The book works well on three fronts. First, it brings the ideas together in an engaging way. The selection of papers that are discussed is well-judged, and the commentaries are clear. Second, it shows a practical mindset. Little always asks how the methods work with real data. Third, the writing is inviting. An informal, first-person voice gives the chapters a seminar feel that will work well for graduate courses and reading groups. A few choices will invite debate. Given the book's scope, some classical figures and topics inevitably receive less attention than some readers would prefer. For example, the book does not mention Herbert Robbins, whose pioneering work (Robbins, 1956) laid the foundation for empirical Bayes estimation and whom Efron (2003) explicitly credited with originating the approach; Efron's own contributions form the central papers discussed in Chapter 7 on empirical Bayes methods. The selection of topics and historical figures, however, is generally coherent.

Because this review is for *JEBS*, I sketch some links between the topics in the book and some themes of psychometric and educational statistics research.

Foundations and latent variables. In latent variable models, the choice between marginal estimation (treating person parameters as random effects), joint estimation (treating them as fixed effects), and Bayesian inference is not only a question of inferential philosophy (see, e.g., Chen et al., 2025); it fixes how we conceptualize the person parameter. That choice carries practical consequences for how scores are interpreted, how uncertainty is reported, and how adaptive designs are built. Thus, in educational and behavioral statistics, debates over frequentist versus Bayesian methods are often coupled with questions about the philosophy of the research domain. In psychometrics, for example, statistical choices are often tied to psychological theory. Haslbeck et al. (2022) is just one example.

Fiducial ideas. Fiducial methods have seen renewed but still limited use in psychometrics (e.g., Liu and Hannig (2016, 2017)). For many common models in behavioral and educational statistics, the area remains largely unexplored. Little's chapter summarizes the arguments for and against fiducial reasoning; for *JEBS* readers, this highlights an area where fiducial

approaches might be compared with established frequentist and Bayesian methods.

Multiple testing. Part II includes a discussion of multiple comparisons and false discovery rate control, a topic that has long been of interest to *JEBS* readers. In fact, some of the early work on multiple testing in which applications to educational data are presented appeared in this journal (Williams et al., 1999; Benjamini and Hochberg, 2000; Thissen et al., 2002; Olejnik et al., 1997). These papers demonstrated the advantages of false discovery rate control over traditional familywise error rate methods, illustrated efficient implementation approaches, and compared the power of various multiple testing procedures, using examples drawn from educational and behavioral research.

Randomized and quasi-experimental designs. Part III's treatment of randomization in clinical trials and propensity score methods for observational studies connects directly to methodological work that has appeared in *JEBS* on experimental and quasi-experimental designs for educational policy evaluation. This includes research on design trade-offs when contamination is a concern (Rhoads, 2011), the role of covariate measurement quality in propensity score adjustments (Steiner et al., 2011), and frameworks for designing multisite trials that examine treatment effect heterogeneity (Dong et al., 2021). These contributions illustrate how the foundational ideas discussed in Little's final chapters have been extended to address challenges in policy evaluation for educational settings.

Sum scores vs optimal scores. In my opinion, there is a strong pragmatic tradition in the field of educational and behavioral statistics, but the historical debates in this book show the value of stating positions clearly. As a small example, consider the debate on the use of sum scores as a way to score respondents. In a recent article, Sijtsma et al. (2024) argue that the sum score is *psychometrics' greatest accomplishment*. In Ramsay and Wiberg

(2017), the authors argue for a strategy to replace sum scores, and state that *a differential weighting of items where the weights depend on ability level as well as items can markedly improve the estimation of ability.* Such contrasting views illustrate how statistical choices reflect deeper perspectives on measurement, not just technical preferences.

Conclusion

Seminal Ideas and Controversies in Statistics offers a clear and thoughtful account of debates that have shaped modern statistics. I recommend it to anyone interested in the development of the field, from researchers and graduate students to practitioners looking for perspective, as a way to deepen understanding of how today's methods have emerged from longstanding debates, and as a reminder of the value of continuing such discussions.

References

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83.

Chen, Y., Li, X., Liu, J., and Ying, Z. (2025). Item response theory—a statistical framework for educational and psychological measurement. *Statistical Science*, 40(2):167–194.

Dong, N., Kelcey, B., and Spybrook, J. (2021). Design considerations in multisite randomized trials probing moderated treatment effects. *Journal of Educational and Behavioral Statistics*, 46(5):527–559.

Efron, B. (2003). Robbins, empirical bayes and microarrays. *Annals of Statistics*, 31(2):366–378.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, 3(3):445–450.

Haslbeck, J., Ryan, O., Robinaugh, D. J., Waldorp, L. J., and Borsboom, D. (2022). Modeling psychopathology: From data models to formal theories. *Psychological Methods*, 27(6):930.

Holland, P. W. (2013). Comments on van der Linden's critique and proposal for equating. *Journal of Educational Measurement*, 50(3):286–294.

Liu, Y. and Hannig, J. (2016). Generalized fiducial inference for binary logistic item response models. *Psychometrika*, 81(2):290–324.

Liu, Y. and Hannig, J. (2017). Generalized fiducial inference for logistic graded response models. *Psychometrika*, 82(4):1097–1125.

Olejnik, S., Li, J., Supattathum, S., and Huberty, C. J. (1997). Multiple testing and statistical power with modified bonferroni procedures. *Journal of Educational and Behavioral Statistics*, 22(4):389–406.

Ramsay, J. O. and Wiberg, M. (2017). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics*, 42(3):282–307.

Rhoads, C. H. (2011). The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics*, 36(1):76–104.

Robbins, H. (1956). An empirical bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pages 157–163. University of California Press.

Sijtsma, K., Ellis, J. L., and Borsboom, D. (2024). Recognize the value of the sum score, psychometrics' greatest accomplishment. *Psychometrika*, 89(1):84–117.

Steiner, P. M., Cook, T. D., and Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36(2):213–236.

Thissen, D., Steinberg, L., and Kuang, D. (2002). Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral statistics*, 27(1):77–83.

van der Linden, W. J. (2013). Some conceptual issues in observed-score equating. *Journal of Educational Measurement*, 50(3):249–285.

Wasserman, L. (2008). Comment on article by Gelman. *Bayesian Analysis*, 3(3):463–466.

Williams, V. S., Jones, L. V., and Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, 24(1):42–69.