Behaviormetrika (2022) 49:165–167
https://doi.org/10.1007/s41237-022-00175-x

PREFACE

Special issue: shadow-test approach to adaptive testing

Wim J. van der Linden[1] · **Maomi Ueno**[2]

Published online: 30 June 2022

In early research on adaptive testing, the common point of view was that of a computer algorithm picking one item at a time at an ability parameter updated after each new response by the test taker. The typical research questions, usually answered with the help of a computer simulation study, addressed such problems as the best initial ability estimate to be fed into the algorithm, the criterion for selecting the next item, the update of the test taker's ability parameter, and the stopping rule that should be used.

The first experiences with real-world adaptive testing programs quickly led to a much larger series of questions though, some still purely statistical but others of a more practical nature. One of the first questions was how to minimize the danger of security breaches due to overexposure of the items with the best statistical parameters in the item pool. Questions that followed were raised for such problems as the maintenance of a content blueprint across administrations of the test, the enforcement of the same degree of test speededness for the test takers no matter the selection of items offered to them, the presence of items organized as sets around a common stimulus, and the possibility of calibrating new items online during adaptive testing.

As these and other questions emerged, they were typically addressed one at a time. Consequently, despite their technical sophistication, not many of the solutions have reached the practice of adaptive testing, the reason being the existence of powerful tradeoffs between nearly every parameter in the design of an adaptive testing program. Because of these tradeoffs, solutions that behave well in isolation are bound to have unintended consequences when used in operational testing. For instance, modifications of the algorithm to control the risk of item compromise face the dilemma between reduced exposure rates and loss of accuracy of the test takers' scores. However, any solution to this dilemma should also maintain the content blueprint for all test takers. And for the solution to be acceptable, it should also resolve the problem of differential speededness inherent in

✉ Wim J. van der Linden
wjvdlinden@outlook.com

[1] University of Twente, Enschede, The Netherlands

[2] University of Electro-Communications, Tokyo, Japan

adaptive testing. Therefore, rather than a series of local solutions, what is needed is a global solution that allows us to manage all parameters of a testing program simultaneously, preferably with software fast enough to serve potentially large numbers of test takers in real time. The shadow test offers such a solution.

The first article by van der Linden (2022) reviews the shadow-test approach. It begins with a formal characterization of adaptive testing as a severely constrained discrete optimization problem with a solution that should be obtained sequentially without any backtracking to earlier items. The shadow-test approach finds the solution combining mixed integer programming (MIP) modeling, sequential Bayesian updating of the parameters of the program, and the use of statistical optimal design theory. It is shown how the approach can be used to manage adaptive testing programs through the choice of a set of linear constraint meeting each of the statistical and practical specifications in force for the testing program.

The second article by Choi et al. (2022) highlights **TestDesign**, an R package based on a generalization of the shadow test approach that can be used to run testing program with exactly the same content specifications but any level of adaptation, from a fixed-form testing format, through formats as linear on the fly, multistage with fixed subtests, multistage with adaptive routing, all the way to item-level adaptive testing. The package offers simple menus to enter each of the constraints necessary to model the content and/or practical test specifications and offers links to a variety of MIP solvers that can be used to assemble the test. A large variety of informative graphical displays is available to evaluate administrations of the test. Worked empirical examples show how to run the software in a variety of applications.

One of the most complicated cases of adaptive testing arises when the item pool contains items organized as sets around a common stimulus. Typical examples are reading comprehension tests gauging the test takers' understanding of a given text and science tests with questions about a display of data found in a physical experiment. The problem is complicated because, in addition to the types of constraints commonly met for tests with discrete items, we now face such requirements as the presence of additional constraint levels for the stimuli, the necessity to deal both with between-set and within-set adaptations of the item selection, as well as possible order restrictions between items within sets. The article by Choi and Lim (2022) shows results from a case study in which the shadow-test approach met each of these requirements for testing formats with any level of adaptation. The output shows results without violation of any of the constraints for any of the test takers but ability estimation still optimal given the set of constraints.

The appropriate way to deal with unknown parameters in statistical models adopted to run a testing program is through Bayesian statistics. For adaptive testing, the approach should be sequential Bayesian with the posterior distribution from the previous update serving as prior for the next. As was already known, the approach simplifies enormously when using Gibbs sampling with resampling of the posterior distributions of all nuisance parameters in the testing model but a Metropolis–Hastings step for the intentional parameters at each iteration. The last article by Niu and Choi (2022) addresses the choice of the proposal density for the sampler and present

implementations with running times for each of the updates of the ability parameter in the range of only 8–17 ms for adaptive tests from a pool of 210 items.

References

Choi SW, Lim S (2022) Adaptive test assembly with a mix of set-based and discrete items. Behaviormetrika. https://doi.org/10.1007/s41237-021-00148-6

Choi SW, Lim S, van der Linden WJ (2022) TestDesign: an optimal design approach to constructing fixed and adaptive tests in R. Behaviormetrika. https://doi.org/10.1007/s41237-021-00145-9

Niu L, Choi SW (2022) More efficient fully Bayesian adaptive testing with a revised proposal distribution. Behaviormetrika. https://doi.org/10.1007/s41237-021-00156-6

van der Linden WJ (2022) Review of the shadow-test approach to adaptive testing. Behaviormetrika. https://doi.org/10.1007/s41237-021-00150-y