TestGardener: A Program for Optimal Scoring and Graphical Analysis

Juan Li, James O. Ramsay and Marie Wiberg

Abstract The aim of this paper is to demonstrate how to use TestGardener to analyze testing data with various item types and explain some main displays. TestGardener is a software designed to aid the development, evaluation, and use of multiple choice examinations, psychological scales, questionnaires, and similar types of data. This software implements the optimal scoring of binary and multi-option items, and uses spline smoothing to obtain item characteristics curves (ICCs) that better fit the real data. Using TestGardner does not require any programming skill or formal statistical knowledge, which will make optimal scoring and item response theory more approachable for test analysts, test developers, researchers, and general public.

Keywords Item response theory · Graphical analysis software · Optimal scoring · Spline smoothing

1 Introduction

TestGardener is the successor of TestGraf, and both softwares are designed to aid the development, evaluation, and use of multiple-choice examinations, psychological scales, questionnaires, and similar types of data. TestGraf was developed by James Ramsay (1995) and has been widely used as an analysis and/or teaching tool of nonparametric item response theory (IRT) in fields like education (Liane, 1995; Nering & Ostini, 2010), psychology (Lévesque et al., 2017; Sachs et al., 2003), medicine (Gomez, 2007; Luciano et al., 2010), and business (Laroche et al., 1999). Users who are familiar with TestGraf can still choose to use its algorithms such as item correct score and kernel smoothing within the TestGardener framework. But

J. Li (✉)
Department of Mathematics and Statistics, McGill University, Montreal, Canada
e-mail: lijuan.640@gmail.com

J. O. Ramsay
Department of Psychology, McGill University, Montreal, Canada

M. Wiberg
Department of Statistics, USBE, Umeå University, Umeå, Sweden

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_8

this paper will focus on the new features (spline smoothing and optimal scoring) and displays that are included in TestGardener.

When we analyze and evaluate real-world testing data, a known problem with parametric IRT is the inability to model all items in a test accurately, even in carefully developed large-scale tests. Using spline smoothing, TestGardener can provide estimated item characteristic curves (ICC) with high precision, even for ill-behaved items. TestGardener also implements optimal scores, which consider the interaction between the test-takers' performance and the sensitivity of the items.

With the user-friendly interface and self-explanatory displays, TestGardener is designed for users with various backgrounds, with or without knowledge in IRT, statistics, and programming. Psychometricians, researchers, test developers, and teachers can easily upload their data, and have the analysis results displayed in diagrams.

TestGardener is relatively fast when analyzing real-world testing data. A sample of 54,033 test takers response data who took the quantitative part of the Swedish Scholastic Assessment Test (SweSAT) is used to demonstrate TestGardener. The SweSAT is a multiple-choice college admissions test, with a verbal and a quantitative part, each containing 80 items. The whole analysis of this 54,033*80 multi-choice data, including reading and writing files, takes about five minutes using a laptop with intel i7 core.

The next section briefly introduces the algorithms of spline smoothing and optimal scoring, which are implemented in TestGardener. The following section provides a short demo of using this software and describes some of the main displays. This paper ends with a short discussion about different versions of TestGardener, new features that may be implemented in later version, and some closing remarks.

2 Theories Behind TestGardener

The real-world testing data rarely meets all the assumptions made in the parametric IRT model. Taking one administration of SweSAT (quantitative part) as an example, the distribution of sum scores is much more skewed than the normal distribution (Fig. 1), indicating that it was a difficult test.

Furthermore, the highlighted ICCs show the ill-behaviors of some items: some items have almost linear ICCs (see highlighted curve in Fig. 2), which means that these items are not very discriminating at any ability level. There are also items with plateaus for a certain score range (Fig. 2); it means these items have no sensitivity for test takers in these ranges. It's probably because test takers with certain level of related knowledge can rule out some of the distractors and choose among the rest options. ICCs of these ill-behaved items can be difficult to estimate using parametric IRT. But TestGardener, using spline smoothing, can estimate these 80 curves without any problems and in only a few seconds.

It is important that the test scores should estimate the test takers' ability as precisely as possible, since tests and test scores are often used to make decisions about test

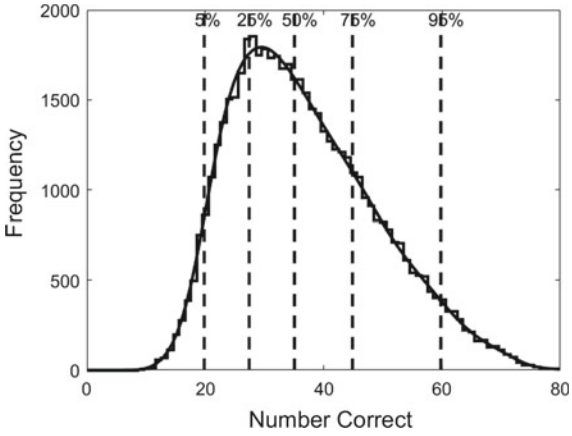

Fig. 1 Distribution of sum scores of one administration of the SweSAT. The histogram indicates the number of test takers within each score range, the black smooth line indicates the smooth density function. The vertical dotted lines are the 5, 25, 50, 75, and 95% quantile lines

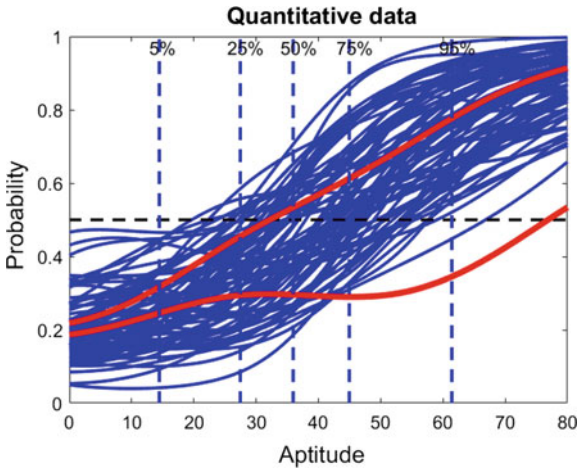

Fig. 2 Estimated ICCs of one administration of the SweSAT, quantitative part. Blue curves are the ICCs of all the 80 items, and red curves highlighted some of the ill-behaved items

takers. Sum score (or number correct score) has been the most commonly used test score because it is easy to interpret and computationally fast. However, sum scores assume that a certain item has constant sensitivity over the entire ability range, which is seldom true. For example, an easy item can have high discrimination power for lower-end test takers, but provides virtually no information about the top students, and vice versa. Optimal scoring, first proposed by Ramsay and Wiberg (2017a), considers the interaction between performance/ability-level and item sensitivity and provides more precise estimation of the test takers' ability. In their 2017 paper, Ramsay and Wiberg only considered the binary response (0/1); but with the extra information of which (wrong) option has been chosen, we can have even more precise estimation of the ability θ because sometimes some wrong options are more wrong than others.

Let S_j denote sum scores and let $P_i(\theta)$ be the probability that the test taker j with ability level θ answer an item i correctly, $i = 1, \ldots, n$; $j = 1, \ldots, N$. The estimate

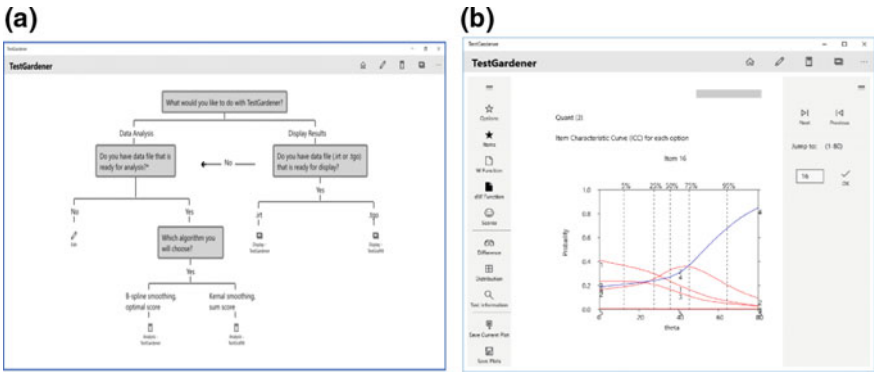

Fig. 3 The opening page (**a**) and display page (**b**) of TestGardener

of optimal scores focuses on estimating the more convenient choice $W_i(\theta)$ as it will facilitate the estimations (Ramsay & Wiberg 2017a). $W_i(\theta)$ is the log-odds of $P_i(\theta)$, which can be defined in terms of $P_i(\theta)$ as

$$W_i(\theta) = \log\left(\frac{P_i(\theta)}{1 - P_i(\theta)}\right). \tag{1}$$

If U_{ij} is test taker j response to item i and if either $P_i(\theta)$ or their counterparts $W_i(\theta)$ are known or we can condition on estimates on them, then the optimal θ associated with the negative log likelihood satisfies the equation

$$\sum_i^n \sum_m^{M_i} [U_{ji,m} - P_{i,m}(\theta_j)] \frac{\mathrm{d}W_{i,m}}{\mathrm{d}\theta_j} = 0, \tag{2}$$

where $m = 1, \ldots, M_i$ and M_i is the number of options of item i. More details about optimal scoring can be found in Ramsay and Wiberg (2017a, 2017b) and Wiberg, Ramsay, and Li (2018). More papers about optimal scoring of multi-choice items and scale items are currently under preparation.

3 A Short Demo of TestGardener

Using TestGardener requires no knowledge of programming; users can simply upload their data (in format described in the manual) and have it analyzed. A result file in .irt format will be generated by the software; it stores all the analysis results and will be used to generate the graphical displays. Figure 3a shows the opening page of TestGardener. By following the flow chart, users should be able to find the appropriate function. Users have the option to change values of several important parameters; but for most users, they are recommended to run the analysis with default values.

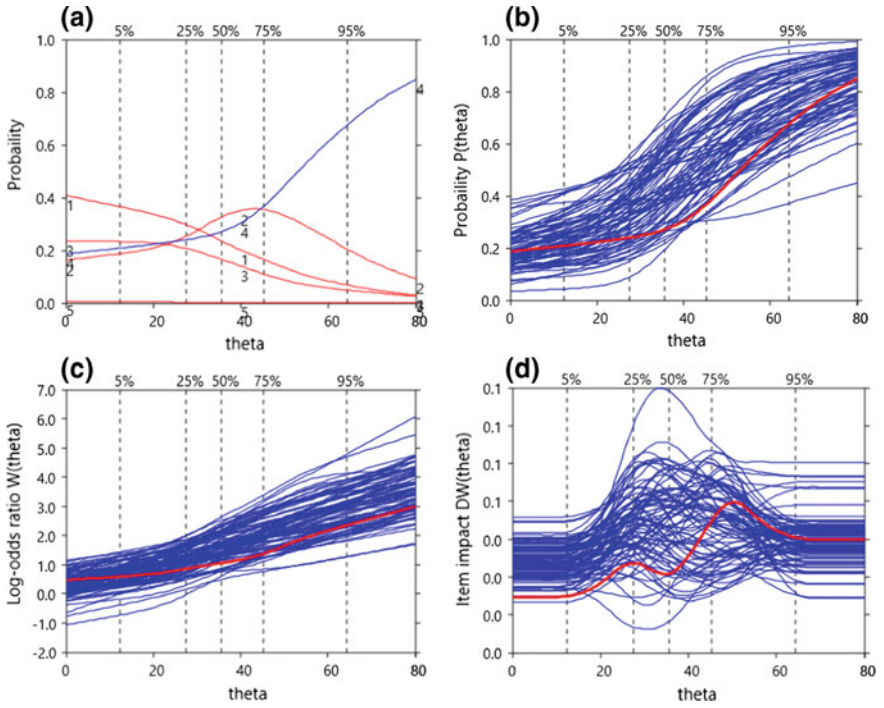

Fig. 4 ICC and other displays indicating performance of item 16. **a** ICCs of all options, blue curve indicates the right option, and red curves represent wrong options. Display **b–d** are for the right option only; $P_{16}(\theta)$, $W_{16}(\theta)$, and the derivative of $W_{16}(\theta)$ respectively. The blue curves are the corresponding curves of all the items, in which the curves of item 16 is highlighted with red

Using the .irt file, users can review the performance of item, test-taker, and test in various displays. The left panel in Fig. 3b lists the names of different displays, which will be introduced briefly below.

Figure 4 shows four displays that represents the performance of an individual item, here we randomly select item 16 as an example. Figure 4a shows the ICCs of all the options, where the right option is represented by a blue line. The indices associating with each curve indicate the corresponding option, so test developer or analyst can have more detailed evaluation of each option. For example, in item 16, option 2 seems quite distracting for test-takers in the middle to upper range. In fact, even for the top students, there is still around 10% probability that they may choose option 2. Figure 4b–d illustrate the probability ($P_{16}(\theta)$), the log-odds ratio ($W_{16}(\theta)$), and derivative of $W_{16}(\theta)$ ($dW_{16}(\theta)$) respectively. $W_i(\theta)$ and $dW_i(\theta)$ curves illustrate the items' sensitivity at each score value and are especially important for the process of optimal scoring. With the corresponding curves of all items (blue curves) in the background, users can have a more intuitive impression of how this item performed comparing with other items.

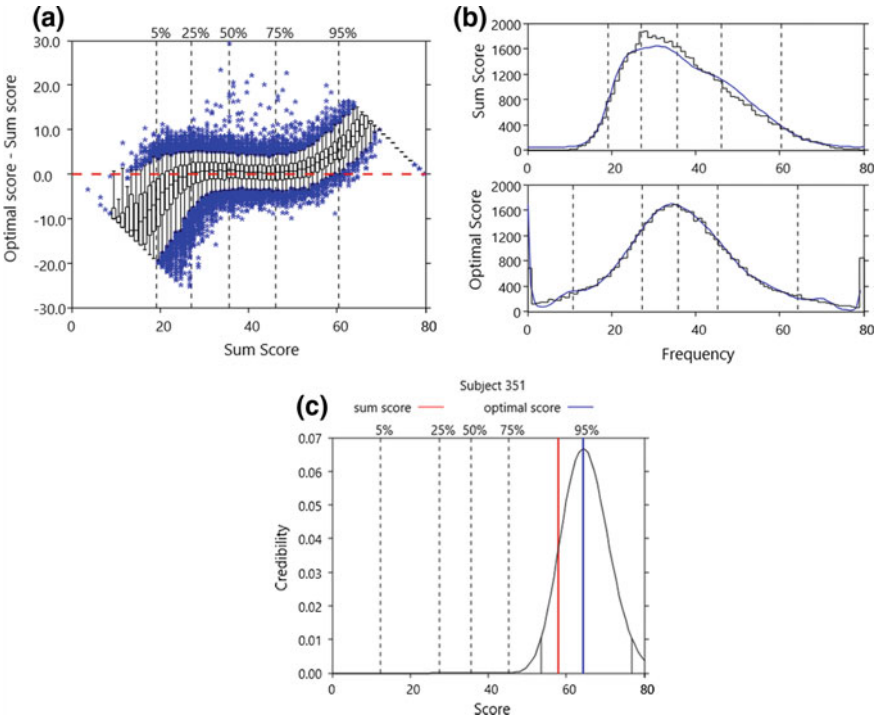

Fig. 5 Displays about the comparison between sum score and optimal score: **a** box plots of the difference between optimal score and sum score; **b** two-panel plot of the distributions of sum score and optimal score respectively; and **c** score credibility plot of subject 351: red and blue vertical lines indicate the sum score and optimal score respectively; black curve shows the likelihood (credibility) of the score, and the two black vertical lines indicate the 95% confidence interval

Figure 5 contains three displays that show the comparison between optimal score and sum score. Since the sum scores are integers, we can plot the difference between optimal score and sum score using box plots. Figure 5a shows that the differences in the middle range are distributed around zero, while for the lower and upper end, the differences are mostly negative and positive respectively. For lower end, optimal scores are corrected for guessing; while for upper end, optimal scores eliminate the influence of ill-behaved items. Figure 5b shows the distribution of sum score and optimal score, with quantile lines changing accordingly. It conveys the same information as the boxplot but from another perspective. Figure 5c shows the comparison at the individual level using the likelihood curve. Optimal score is always at the peak of the likelihood curve, hence optimal.

4 Discussion

TestGardener currently has two versions: one stand-alone version for windows system and one web-based version that can be used on any major browsers. These two versions share almost the same features, but currently, the stand-alone version has more ability to edit and prepare the data file. The web-based version is newly developed for users on other operating systems or someone who wants to try some of the features before downloading the software. It also serves as the teaching platform for optimal scoring and even item response theory, by including pages for software manual, theories, and resources. Both versions of TestGardener are still under development, although a beta version dedicating to the workshop held in Umea, Sweden this August has been published. Readers who are interested in TestGardener are welcome to try the web-based version on http://testgardener.azurewebsites.net/. But please note that this version is premature and not very stable.

The formally released TestGardener (both versions) are expected to be even faster, with more display options. For example, users may be able to choose the options that plot the confidence interval and data points on the ICCs. Plots in Fig. 4 are currently displayed separately, and we plan to implement this four-panel plot like Fig. 4 in the later version of TestGardener.

Acknowledgements This research was funded by the Swedish Research Council (grant. 2014-578).

References

Gomez, R. (2007). Australian parent and teacher ratings of the DSM-IV ADHD symptoms differential symptom functioning and parent-teacher agreement and differences. *Journal of Attention Disorders*. *11*(1), 17–27.

Laroche, M., Chankon, K., & Tomiuk, M. (1999). Irt-based item level analysis: an additional diagnostic tool for scale purification. In J. E. Arnould, L. M. Scott (Eds.) *Advances in consumer research* (Vol 26, pp. 141–149). Provo, UT: Association for Consumer Research.

Lévesque, D., Sévigny, S., Giroux, I., & Jacques, C. (2017). Gambling-related cognition scale (GRCS): Are skills-based games at a disadvantage? *Psychology of Addictive Behaviors, 31*(6), 647–654.

Liane, P. (1995). A comparison of item parameter estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes. The University of Ottawa, thesis.

Luciano, J., Ayuso-Mateos, J., Aguado, J., Fernandez, A., Serrano-Blance, A., Roca, M., et al. (2010). The 12-item world health organization disability assessment schedule II (WHO-DAS II): A nonparametric item response analysis. *BMC Medical Research Methodology, 2010*(10), 45.

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York: Taylor and Francis.

Ramsay, J. O. (1995). *TestGraf—a program for the graphical analysis of multiple choice test and questionnaire data [computer software]*. Montreal: McGill University.

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

Ramsay, J. O. & Wiberg, M. (2017b). *Breaking through the sum score barrier.* (pp. 151–158). Paper presented at the International Meeting of the Psychometric Society, Asheville: NC, July 11–15.
Sachs, J., Law, Y., & Chan, C. K. (2003). A nonparametric item analysis of a selected item subset of the learning process questionnaire. *British Journal of Educational Psychology 73*(3), 395–423.
Wiberg, M., Ramsay, J. O., & Li, J. (2018). Optimal scores as an alternative to sum scores. In: M. Wiberg, S. Culpepper, R. Janssen, J. González, D. Molenaar (eds) *Quantitative Psychology. IMPS 2017*. Springer Proceedings in Mathematics & Statistics, vol 233. Cham: Springer.