<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *Journal of educational and behavioral statistics*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Ramsay, J., Wiberg, M., Li, J. (2020)

Full Information Optimal Scoring

Journal of educational and behavioral statistics, 45(3): 297-315

<https://doi.org/10.3102/1076998619885636>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-165757>

Full Information Optimal Scoring

J. O. Ramsay

Department of Psychology, McGill University

J. Li

Ottawa Hospital Research Institute

M. Wiberg

Department of Statistics, USBE, Umeå University

20 June 2019

The authors gratefully acknowledge the support to this research provided by Natural Science and Engineering Research Council of Canada Discovery Grant 320-2012-RGPIN to J. O. Ramsay and the Swedish Research Council grant 2014-578 to M. Wiberg.

This paper is published. Please cite as: Ramsay, J.O, Li, J., & Wiberg, M. (2020). Full information optimal scoring. *Journal of Educational and Behavioral Statistics*, 45(3), 297-315. <http://dx.doi.org/10.3102/1076998619885636>

Abstract

Ramsay and Wiberg (2017) used a new version of item response theory that represents test performance over non-negative closed intervals such as $[0,100]$ or $[0,n]$ and demonstrated that optimal scoring of binary test data yielded substantial improvements in point-wise root-mean-squared error and bias over number right or sum scoring. We extend these results by showing that optimal scoring of the full information in option choices produces about as much further improvement in these measures of score performance as was achieved by going from sum scoring to optimal binary scoring. We also offer a new fast algorithm for estimating option response functions and optimal scoring of each examinee's data. Finally we propose the use of the concept surprisal for evolution of option choices over the score interval and a distinction between score index values and test scores.

Key Words: option characteristic curves, option surprisal function, option sensitivity function, sum scoring, optimal scoring, four-stage test analysis cycle

1 Introduction and Overview

The goal of item response theory (IRT), as with most statistical models, is to provide better information to those who use the data to characterize examinee performance and to make decisions on the basis of this characterization. Here the characterization is usually a real number θ on a line segment from which ranks and binary decisions can be defined. A variety of powerful strategies have been explored for either estimating item response functions and examinee scores simultaneously, or proceeding in a two-stage fashion by first identifying item features and then estimating scores conditioned

on these feature estimates. We refer to the scores that are defined either way as *optimal* in the sense that they minimize some loss function such as negative log likelihood that statisticians take seriously. Of course that leaves open the possibility that some “optimal” scores are better than others, perhaps by using a model that is better at characterizing the data.

The great majority of examinations, not excluding those developed by testing agencies, are scored by counting the number of correct answers or otherwise summing pre-allocated item scores. This scoring strategy is easy to apply and intuitively appealing, but is not optimal in our sense. Many aspects of conventional IRT may have contributed to retention of the sum score, but it is probably the representation of the score on the whole real line that the potential users find the most unpalatable. In addition to a distaste for signed numbers, the use of infinity to represent none and all right answers collides with the fact that the data from which scores are extracted collapse to binary information at the extremes of performance, strongly suggesting for most of us that it would better to use 0 on the left and either 100 or the number of items n on the right. In these aspects the sum score has it right.

The use of the whole real line also severely handicaps the comparison of the efficiency of IRT-based scores and sum scores. The widespread use of correlations in the psychometric literature as measures of agreement cannot reveal whether sum scores may be more inefficient over some performance ranges than over others or characterize poor performance in terms of the two orthogonal dimensions of bias and sampling variance.

But IRT does nevertheless capture two ways in which the sum score falls short. The first of these is the sum score’s failure to take item performance into account. Items can provide useful information over some score ranges but not over others; easy items are useful for terminating educational support and hard items tell who the elites

are, and no item can be highly informative over all performance levels. In short, there is an interaction between item and examinee performance that the sum score ignores.

The second deficiency of sum scores in the context of multiple choice items is the failure to use the information in item options declared by the test developer to be wrong. That is, the reduction of the data to binary form ignores the fact that some wrong choices do indicate at least partial knowledge, and occasionally even that a “wrong” answer can be seen from some perspective as legitimate.

Ramsay and Wiberg (2017) demonstrated the substantial superiority of optimal test scoring over the usual number right or sum score in the context of binary scored right/wrong data. Their primary innovations were (1) to model n item response curves over the same $[0, n]$ or $[0, 100]$ intervals used to report sum scores, (2) to use the log-odds functions $W_i(\theta) = \log\{P_i(\theta)/[1 - P_i(\theta)]\}$, $i = 1, \dots, n$ to represent item characteristics, (3) to employ functional or nonparametric methods for estimating the shapes of these log-odds functions (see also Lee (2007), Ramsay (1997), Ramsay and Silverman (2002), Rossi, Wang and Ramsay (2002), Woods (2006) and Woods and Thissen (2006)) and (4) to compare score performances directly in terms of both bias and root-mean-squared-error (RMSE) as functions of ability θ .

They attributed the better performance of optimal scores to their use of item/examinee interaction effects as opposed to the uniform non-interactive weighting implied by sum scoring. They also highlighted the especially strong negative bias in the sum scores for the most able examinees as well as the over-estimation of ability by sum scores for the weakest examinees. Wiberg, Ramsay and Li (2018a) compared optimal test scoring with sum scores and expected sum scores and found optimal scoring to have the least bias and to be the most efficient in terms of RMSE. Similarly, Wiberg, Ramsay and Li (2018b) compared optimal test scoring with sum scoring and IRT scores and found optimal scoring to have smaller bias and be more efficient in terms of MSE.

The aim of this paper is to examine what additional test scoring performance can be achieved by using the complete option choice information, including missing and spoiled responses. That is, we ask whether knowing wrong option choices might further improve score accuracy by taking into account whether the option chosen reflects at least partial knowledge as opposed to option choices that might reflect pure guessing or other completely uninformed strategies.

We also introduce a number of modifications of the Ramsay and Wiberg (2017) test analysis methodology that can considerably speed up analyses with negligible loss in score accuracy. In particular, we recommend analyzing tests in two stages. The first stage involves the estimation of the full information equivalent of log-odds option characteristic curves using nonparametric data smoothing over test scores, and the second optimally scores each examinee’s data given the the estimated option curves.

The data that we use for both illustration and simulation were the complete option choice records for 53,768 examinees who took both the verbal and quantitative subtests of an administration of the Swedish Scholastic Assessment Test, abbreviated here as the SweSAT. Each subtest was administered in two sessions with 40 items per session. There were five options for the first ten items of each of the verbal subtest and four for the remainder, and items 23 to 28 had five options and the remainder four options for the quantitative subtest. We added to each item an additional “option” to represent items that were either not attempted or had spoiled responses.

Figure 1 displays smooth estimates of the probability density functions $f_V(\theta)$ and $f_Q(\theta)$ for the verbal and quantitative SweSAT subtests and for both optimal and sum scores, respectively. Both subtests are difficult, but the quantitative subtest was more so, with a median score of only 35 out of 80 as opposed to a median of 43 for the verbal subtest. The main contrast between optimal and sum scoring is in terms of the densities at the extreme score ranges. Optimal scoring estimates lower

Table 1: The frequencies of extreme scores among 53,768 examinees for the verbal and quantitative subtests of the SweSAT.

Subtest	Score	Frequency	
		0	80
Verbal	Sum	0	3
	Optimal	42	216
Quantitative	Sum	0	2
	Optimal	47	143

scores than sum scoring for the weakest examinees, but estimates higher scores for the strongest examinees. Table 1 contains the numbers of extreme scores for sum and optimal scoring. These results are consistent with the bias results found by Ramsay and Wiberg (2017), where optimal scoring was found to correct the positive bias in sum scores due to guessing for weaker examinees and the negative sum score bias due to defective items for the most proficient examinees.

Section 2 lays out the model for multinomial information tests as opposed to only right/wrong binary information. Section 3 provides technical details concerning the data smoothing process for estimating option characteristic functions. Section 4 then provides optional scoring details including possible smoothing of the score distribution toward a target distribution. Section 5 presents simulation results that compare the performance of sum scores, optimal scoring with binary data and optimal scoring with full data.

Figure 1: The staircase line is a histogram of optimal scores with bins defined by the integers from 0 to 80. The solid line is a nonparametric estimate of the probability density of these scores, and the dashed line is the corresponding density for the sum scores. The vertical dashed lines are the respective 5%, 25%, 50%, 75% and 95% quantiles of their respective optimal score distributions.

2 The Full Option Information Model and Option Surprisal Functions

The number of options presented to examinees for the SweSAT subtests varied between four and five, but the data also included indicators that items that were not responded to, as well as whether the response was otherwise invalid. Since the probabilities must sum to one, we added a single additional option to each item indicating that either a non- or a spoiled response was coded. The total numbers of options were 420 and 412 for the verbal and quantitative subtests, respectively. Figure 2 displays the number of non-responses and spoiled responses for both subtests and both sessions. We see that there are a large number of non-responses at the end of the first session (item 40), but rather fewer for the second session (item 80). We also suspect that the non-responses for items 33-40 for the verbal are primarily due to the examinee running out of time, whereas the non-response pattern for the quantitative items suggests that examinees bypass of especially difficult quantitative items and often do not have time to return to them.

2.1 The Option Surprisal Function

Let $P_{im}(\theta); i = 1, \dots, n; m = 1, \dots, M_i$ indicate the probability of an examinee at ability level θ choosing option m for item i and let $W_{im}(\theta) = -\log P_{im}(\theta)$ so that $P_{im}(\theta) = \exp[-W_{im}(\theta)]$. where $W_{im}(\theta)$ is a non-negative function that is unbounded below and constrained to satisfy the normalization constraint $\sum_m \exp[-W_{im}(\theta)] = 1$ for all i and θ .

The quantity $W = -\log_2 P$ is called *surprisal* in information theory, and is widely used in statistical modelling within physics, chemistry and biomedical engineering

Figure 2: The percentages of either non-responses or invalid responses for each item for the two SweSAT subtests.

as a replacement for the concept of probability (Tribus, 1961). Surprisal enjoys the advantages of having a lower limit of 0, meaning certainty, and of possessing the operations of addition and differencing. That is, it behaves like the magnitudes that we experience and that are the fundamental physical quantities. Its unit is the bit, but when the natural logarithm is used, the unit of surprisal is e and is called the *ebit*. Surprisal zero corresponds to probability one. Intuitively, we tend to think of a magnitude of zero as a baseline, and therefore not especially interesting. A quantity that increases in size tends to attract our attention, which is not the case for probability.

We prefer defining the option response function by surprisal since $W_{im}(\theta)$ is easier to fit by simple models because it does not possess the upper horizontal asymptote that makes $P_{im}(\theta)$ difficult to approximate directly. More importantly, we will see below that surprisal plays a natural role in defining an optimal score.

Figure 3 displays the shapes of $P_{im}(\theta)$ and $W_{im}(\theta)$ for the 80 correct options for the two subtests. Plotting a lot of curves hides a great deal of detail, but we notice that many probability curves fail to approach one on the right, and that there are a few that are especially deficient. We also note a general tendency for these curves to slope linearly upward over most of the score range, but rise more sharply for the top 5% of the scores. But both sets of curve shapes have a few deviant cases that we need to explain in more detail.

2.2 The Score Index versus the Test Score

In our view, the examinee performance parameter θ has been widely misinterpreted in the psychometric literature via the term *latent variable*. It is a continuous index rather than a *measure* in this model since it only defines position along a continuum.

Figure 3: The top panels display the option curves $P_{im}(\theta)$ for the correct options for the two SweSAT subtests, and the bottom two panels display their corresponding surprisal versions $W_{im}(\theta)$.

Any strictly monotonic transformation of θ would serve equally well as an index, just as any strictly increasing discrete set of numbers would serve to represent the rank of a set of objects. That is, if we redefine the value representing performance as $\xi = h(\theta)$ where h is a strictly increasing and smooth function of θ , that we can always redefine the option probability function so that $P_{im}^*(\xi) = P_{im}(\theta)$, and similarly for surprisal $W_{im}(\theta)$. Consequently, algebraic operations such as summing and differencing θ values have no meaning, except perhaps as approximations over very small neighbourhoods within highly smooth curves W_{im} . What is essential about θ is the continuity that it implies for the evolution of functions like $P_{im}(\theta)$ and $W_{im}(\theta)$. We therefore prefer the term *score index* for θ .

On the other hand, a test score, which we indicate as $\mu(\theta)$, is a quantity that one does not want to vary over alternative choices of score index systems. Moreover, test scores may depend on pre-assigned numerical weights or scores S_{im} assigned to each possible choice. The multiple choice exam situation where $S_{im} = 1$ when the chosen option is correct and zero otherwise is a special case, and does not apply to the vast number of rating scales for which item response theory is also useful. But since the probability value does not vary when score index system changes, we define the test score as

$$\mu(\theta) = \sum_i^n \sum_m^{M_i} S_{im} P_{im}(\theta). \quad (1)$$

Optimal scoring involves finding the score index θ that defines the best fit to the choice data at hand, and then the subsequent application of the test score formula to provide an optimal test score estimate. In this sense, the *observed test score* $\sum_i^n \sum_m^{M_i} S_{im} U_{im}$ is only an estimate of the test score; and, although simple to compute, is not necessarily efficient as an estimate.

There is one choice of score index, however, that has an additive structure and that

might serve also as a test score. As either the probability or surprisal option functions evolve over performance, they jointly trace out a space curve in the hypercube with a dimensionality equal to the total number of options and side lengths equal to the size of the score index interval. Distance along this curve, called *arc length*, and defined, in the case of surprisal, by

$$L(\theta) = \int_0^\theta \sqrt{\sum_i^n \sum_m^{M_i} \left[\frac{dW_{im}}{du} \right]^2} du, \quad (2)$$

is a measure in the strict sense of the effort required to move along the curve from its origin to position θ along the curve. It is also invariant with respect to other choices of score index. Arc length can be added and differenced and has origin 0, and therefore seems to be exactly what item response theorists have been seeking.

2.3 The Option Sensitivity Function

The negative log likelihood with respect to ability parameter θ condition on the option surprisal curves is

$$H(\theta) = \sum_{i=1}^n \sum_m^{M_i} U_{im} W_{im}(\theta) \quad (3)$$

where the data values $U_{im}, m = 1, \dots, M_i$ are indicator variables that are all zero except for the value one for the correct option. We note that the negative log likelihood is itself essentially a surprisal, but with the scalar infinitesimal term $\log(nd\theta)$ being dropped. The optimal score solves the equation

$$\frac{dH}{d\theta} = \sum_{i=1}^n \sum_m^{M_i} U_{im} \frac{dW_{im}}{d\theta} = 0. \quad (4)$$

The surprisal derivative value $dW_{im}/d\theta$ acts as weighting coefficient that defines

the importance of the term in the double summation of items and options within examinees. We like to refer to the option surprisal derivative curves as *option sensitivity curves*. When the slope of an option's W -function is strongly positive, the value of that residual contributes strongly in a positive sense to the two sums defining the H -gradient, and when strongly negative, it strongly subtracts the residual. But when a W -function is relatively flat, the residual is down-weighted and therefore is treated as relatively uninformative with respect to defining the optimal score θ . In this way $dW_{im}/d\theta$ models the *interaction* between option slope and ability at the score location θ .

Figure 4 shows the estimated W -slope functions $dW_{im}/d\theta$ for the right answer options for the two SweSAT subtests. We see in the verbal curves a cluster of items that peak for the lowest 25% of examinees as well as a number of right answer options that have high slopes for both the bottom and the top 5%. The quantitative subtest exhibits two clear clusters of peaks, one for the bottom 25% and another for the top 25% as well as a number of items that are informative for the top 5%. The fanning out of these curves over the bottom and top 5% intervals is a consequence of the small amount data available for estimating curve shapes over these intervals.

Figure 4: The option sensitivity functions $dW_{im}/d\theta$ for the surprisal curves in Figure 3.

Figure 5: The four-step cycle for estimating option response functions, which is initialized by sum scores.

3 A Cycled Four-stage Estimation Procedure

Although Ramsay and Wiberg (2017) employed a joint estimation scheme for estimating both option curves and score values, we now prefer a four-step process in which option functions are first estimated, and then score values computed given these estimated option curves. The result was a far faster analysis process whose accuracy based on our simulations was negligibly different from using their joint optimization procedure. The overall structure of the procedure is captured by the diagram in Figure 5.

3.1 Estimation of the Surprisal Curves $W_{im}(\theta)$ by Smoothing

The estimation of the surprisal curves $W_{im}(\theta)$ presents a number of challenges. Unlike most curve estimation problems, the argument values are far from evenly distributed over the ability range, so that estimates of curve end values will be substantially less accurate than those for middle values. The curves are only mildly nonlinear, but nevertheless we do see, for example, plateaus in several of these curves requiring compound curvature, probably due to the ability to eliminate certain wrong answers that comes with even a limited amount of knowledge. Estimates of the option sensitivity curves are essential for optimal smoothing, as well as for computation of item information functions, and therefore the slopes of the curves must also be sufficiently accurate. This implies that curves must be smooth, and we add to smoothness constraint that the curves should exhibit stable as well as plausible behaviour over the end regions where the data are sparse. Finally, we have a lot of option curves to estimate, so that the high dimensionality of the estimation problem demands that estimates can be computed within a tolerable amount of time.

We therefore opted to estimate the $W_{im}(\theta)$ curves using spline smoothing methods, which are also especially appropriate for curves over closed intervals. The required end-zone slope stability was achieved by using a relatively large number of knots, combined with fairly heavy penalizing of the roughness of the third derivative so that the first derivative is smooth. The results shown in this paper were achieved by using 24 B-spline basis functions and a roughness penalty parameter of 10,000. Ramsay, Hooker and Graves (2009) can be consulted for further details on spline smoothing.

We also found that good results in terms of accuracy, stability and computing speed were achieved by binning the score values. These bins were defined so that the numbers of examinees within each bin were approximately equal, consequently

implying that the bin widths were much larger in the end-zones than in the central regions. Defining the appropriate bin boundaries requires an estimate of the probability density function for the score values, and therefore is the first step in the estimation cycle. We used the functional model described in Ramsay, Hooker and Graves (2009) for this purpose, which was determined by using 11 B-spline basis functions. We defined the number of bins by specifying a target number of examinees within any bin, and the large number of examinees in the SweSAT data suggested that we could aim at about 1000 examinees per bin, for a total of 53 bins. In the case that no examinee within a bin chose a specific option, we replaced the proportion of 0 by 0.0005 to avoid convergence problems in computing the W -values from proportions.

This smoothing process produced estimates of all option surprisal function for either subtest in under a second on a typical laptop or desktop computer. Figure 6 displays the option curves for verbal item 15 and quantitative item 55. We note that values of $P(\theta)$ estimated from 1000 proportions per bin are tightly clustered around the estimated curves.

These two items have been chosen because they are especially harmful for sum scores, but much less so for optimal scoring. In the first panel of Figure 6 displaying item 15 from the verbal subtest, the third option, which is scored as correct, is uninformative because it attracts examinees across the entire score range with little change in probability. On the other hand, option two, although not especially informative either, does exhibit the ogival shape expected of correct options. It is this item that has the deviant shape in the left panels of Figure 3. Optimal scoring deals appropriately with this item because, since the slopes of both of these options are comparatively shallow, the scoring equation (4) will place only a small positive weight on both items two and three, but definitely negative weight both items one and four for central 50% of the examinees.

Figure 6: The data values and the fitted functions $P_{im}(\theta)$ for two items. The options scored as right are solid lines, and incorrect options are dotted. The fifth options are for non-responses or spoiled responses.

The quantitative item 55, on the other hand, is an excellent item for examinees around percentile 25, but a serious problem for sum scoring for examinees in the top quartile where a wrong option is chosen with a nearly constant probability. However, since the slopes are virtually zero in this upper zone, optimal scoring will effectively ignore this item over this region. The question and its optional answers is interesting. “What is the relationship between the function $f(x) = 1/x + x$ and 2, assuming that $x > 2$ ”? The four options are: (1) $f(x) > 2$, (2) $2 > f(x)$, (3) $f(x) = 2$ and (4) Information is insufficient. Option (1) appears to be chosen when $x = 1$ is overlooked as a test case, and is chosen with about the same probability for all performance levels. Strictly speaking, option (4), which was scored as correct, is also wrong, since the equation contains all the information that a bright secondary school student would need to declare that $f(x) \geq 2$. But there is no indication in the option curves that the best students recognized this.

3.2 Optimal Scoring Conditional on the Sensitivity Functions

Once the sensitivity curves are defined, an examinee’s optimal score can be estimated by minimizing (3) with respect to θ . However, the potentially large number of examinees such as we have for the SweSAT subtests requires that the optimization be in a parallel and vectorized fashion as much as possible. Our approach for the purpose was to use a safe-guarded Newton-Raphson algorithm that carried out each optimization step simultaneously for all active values of θ , where an active examinee is one for which either the current derivative (4) is not within the specified tolerance region of zero at the current value, or is at a boundary and the derivative would move θ back into the interior. These iterations terminate when the number of active examinees

either goes to zero, or an upper limit on iterations is reached. Our Matlab code achieves these optimized values for all 53,768 examinees in each SweSAT in about 100 seconds on a desktop computer.

Figure 7 plots the optimal minus sum score for each examinee against the sum score for the two subtests. We see that the score differences for 70% of the students that are between the twenty-fifth and ninety-fifth percentiles are fairly evenly distributed around zero, but that differences can nevertheless be substantial. But the bottom 25% of the students tend to be assigned optimal scores well below their sum scores since optimal scoring effectively corrects for guessing. By contrast, the top 5% of the students tend to gain substantially by optimal scoring, primarily because problem items like those displayed in Figure 6 are severely down-weighted. In fact, the students on the diagonal line of points at the highest optimal score levels are those whose sum scores were corrected to perfect scores. This seems natural since one assumes that the very best students can easily cope with even the most difficult items appropriate at this level. Similarly, it would not be surprising to find that a few hundred students are essentially so weak that their positive sum scores can be attributed entirely to lucky guesses and their optimal score would be zero. Figures 1 and 7 warn us that a sensible target density should take into account the fact that there will be a piling up of optimal scores at one or both boundaries. That is, we may need to define a 0- and n -inflated density for the interval $[0, n]$ or a 0- and 100-inflated density for percentage scores.

Large differences between sum and optimal scores are relatively rare, but are to be particularly found for relatively low sum scores where a substantial amount of guessing may be involved. Figure 8 shows the function $H(\theta)$ and the two score index estimates for two examinees for the quantitative subtest. The left panel shows two minima. The global minimum is at the left boundary, but we feel more comfortable

Figure 7: Boxplots for optimal scores minus sum scores for each sum score values from 0 to 80 for the two SweSAT subtests.

Figure 8: The function $H(\theta)$ for two examinees and the quantitative SweSAT. The solid vertical line is positioned at the optimal score index estimate and the dashed line at the sum score.

with the upper minimum just above 40 since the score index estimate of zero may be due to the inevitable instability of $W(\theta)$ estimates where the data are extremely sparse. Perhaps the same caveat can be raised for the right figure and consequently the sum score value might be preferred. The possibility of such judgement calls can never be ruled out, and test analysis software should provide displays such as these as a convenient option.

A testing agency may choose to define a priori target or template probability density function $f_0(\theta)$ that it wishes to impose on estimated θ values, perhaps as

an strategy in a test equating process. In this case, we can alter the optimization criterion (3) by adding the negative log of the target density multiplied by a non-negative smoothing parameter γ to obtain

$$H_\gamma(\theta) = H(\theta) - \gamma \log f_0(\theta). \quad (5)$$

When $\gamma = 1$, we are effectively using the second term to define the negative log of the prior distribution of θ as what defines the optimal θ . A few additional iterations of the smoothing-scoring cycle will move the resulting distribution of the score to that of the target. As γ increases even further, the distribution of the estimates eventually goes to the target distribution itself.

To summarize, the four-step process requires one or more iterations of the following cycle:

1. Estimate a nonparametric probability density function that accurately describes the distribution of a set of N values of θ .
2. In these data using K bins; and, for each item and each option, compute the proportions of examinees within each bin choosing the option. The bin boundaries were set using the estimate θ -density function.
3. Convert the proportions to surprisal values.
4. Estimate a smooth curve by spline smoothing of surprisal values over bin center values and its derivative.
5. Compute θ score values that are optimal given the smooth sensitivity curves.

These cycles could be iterated until surprisal function values and score values converge, but we found that the results after only two cycle the scores were sufficiently

close to their converged values.

4 Simulated Data Experiments

Our simulation experiments assumed that the sample size in the SweSAT was sufficient to take the estimated surprisal curves as known, and therefore to study the properties of the conditional test score estimates. We simulated 1000 sets of data, for each of which we computed both bias and RMSE's averaged over the samples at the values $W(\theta)$, $\theta = 0, \dots, 80$ for (1) optimal scoring using the full information for all options, (2) for the reduction of this information to binary scoring, and (3) for sum scoring.

Figures 9 and 10 display these results for the verbal and quantitative subtests, respectively. They are quite similar, except that the bias at the highest performances is higher for the more difficult quantitative subtest. For both subtests the bias of the sum scores is negative by as much as seven points for the top examinees and positive by as much as 17 points for weakest examinees, but near zero for the 80% of the scores between percentiles 25 and 95. On the other hand, the bias for both optimal scores is effectively zero for most of the score range and is much smaller than that of the sum scores at the extremes.

MSE is the sum of squared bias and sampling variance, so that the RMSE inevitably reflects bias as well as sampling variance. The full information RMSE dominates that of the binary information scores nearly everywhere for both subtests, and both optimal score RMSEs are far smaller than those for the sum scores at the extreme score ranges. For most of the score range the RMSE for binary scoring is midway between that for the optimal and sum scores of both subtests. At the median score of 35.2 for the quantitative subtest, the RMSE's are 3.3, 3.7 and 4.3 for the full

optimal, binary optimal and sum scores, respectively. Full information scoring results in a 40.8% reduction in MSE over that for sum scoring, suggesting that a sum-scored subtest would have to have over 112 items to have the same RMSE as the optimally scored subtest. The reduction achieved by optimal scoring of the binary version of the data is 25.6%, equivalent to a sum-scored subtest with over 100 items.

The items of the quantitative SweSAT test were grouped into four categories:

1. Mathematical problem solving (24 items)
2. Quantitative comparison (20 items)
3. Quantitative reasoning (12 items)
4. Using information presented in diagrams, tables and maps (24 items)

We wanted to view the performances of the separate test score estimates for each of these question categories in order to examine differences due to the type of question, and also to see whether performances were affected by test length. We generated 5000 random samples for each section. The results are shown for only the median test score in each section, and separately for the binary data and full data analyses.

We now see more bias for the sum score for the first and third item categories, but these values are still relatively small when compared with the root-mean-square error levels (RMSE). The RMSE levels reflect the orders of accuracy shown in the median zone in Figure 10. The variance ratios indicate that roughly the same percentage increases in test length would be required to make the section sum scores as accurate as the full data optimal scores. Not shown are the biases and RMSE's associated with scores in the middle of the top 5% of the examinee scores. There are also much the same as seen in the figures for the total test scores. The two true test score values

Figure 9: The top panel shows the RMSE for full information optimal scoring (solid), binary information optimal scoring (dashed) and sum scoring (dotted) for data simulated using the verbal SweSAT subtest as the population. The bottom panel displays the corresponding biases.

Figure 10: The top panel shows the RMSE for full information optimal scoring (solid), binary information optimal scoring (dashed) and sum scoring (dotted) for data simulated using the quantitative SweSAT subtest as the population. The bottom panel displays the corresponding biases.

Table 2: The performances of the test scores on the four quantitative question categories for the median test score. The three score types are the sum score (SS), binary data optimal score (Bin) and full data optimal score (Full). The final column contains the ratio of the sum score sampling variance to the full optimal score sampling variance, and indicates how much longer the test would have to be to have the sum score sampling variance equal to the optimal score sampling variance.

Section	n	True μ		Bias			RMSE			Variance
		Bin	Full	SS	Bin	Full	SS	Bin	Full	Ratio
1	24	9.31	10.28	-0.27	0.01	0.09	2.34	2.18	2.07	1.27
2	20	9.33	9.28	-0.12	-0.03	0.16	2.16	1.99	1.75	1.52
3	12	5.28	5.17	-0.44	-0.23	-0.17	4.29	3.78	3.71	1.32
4	24	10.28	10.22	-0.07	-0.05	0.06	2.31	2.22	1.99	1.35

for each section differ because these values depend on whether binary data or full data are used in the analyses of the original data within each section.

5 Discussion and Conclusions

We have demonstrated that the seemingly harmless step of simplifying the data on option choices to right/wrong only information seriously degrades the quality of the optimal score estimate. In fact, the optimal full information MSE is about as superior to the binary information MSE as the latter is superior to that of the sum score. A referee has pointed out quite rightly that a full option choice model is in some measure a different model than a right/wrong choice model. But it is also a model that compensates examinees for the kinds of problems displayed for item 55 of the quantitative subtest and item 15 for the verbal subtest. In both cases the two problem option functions are quite flat for the brightest examinees, and will therefore contribute relatively little to defining their optimal score indices. These examinees consequently have a greater chance of being assigned a score index θ at its upper

boundary. It is this that leads to the dramatic improvement in bias for the optimal score over the sum score for score extremes.

Using two cycles of the test analysis algorithm diagrammed in Figure 5 greatly speeds up the computation with almost no loss in efficiency, and permits the optimal scoring and the estimation of option curves within a few minutes for even hundreds of thousands of examinees.

Much of the gain in efficiency by optimal scoring can also be credited to the use of the flexible spline expansions of the surprisal functions, along with a strong penalty on their curvature over score extremes where there is relatively little data. We have demonstrated in Ramsay, Wiberg and Li (2018a) that the use of poorly fitting parametric models such as the Rasch model, the two-parameter and the three-parameter logistic models seriously reduces the efficiency of optimal scoring. These models provide a scalar measure of item discrimination power, but we see in this paper that option quality as well as item quality has to be characterized by an entire function, namely the sensitivity curve. We see little justification for continuing to use these parametric models.

The importance of the surprisal function $W(\theta) = -\log P(\theta)$ arises from how the conditional maximum likelihood estimation equation depends on the option choice model function $P(\theta)$, and is a also useful alternative way of understanding probability or certainty of an option choice. Surprisal has the advantage of behaving the way other magnitudes in science do.

The estimation of choice probabilities and surprisals over extreme test score and score index values is inevitably difficult because of the small number of examinees in these performance zones. We can, however, expect that the choice probabilities for the correct options for multiple choice questions will increase monotonically, as will the initial and final categories in rating scales. Using this prior knowledge can bring

better estimation where it is appropriate, and we are therefore developing methods for monotonic nonparametric curve estimation as a way of achieving further improvement in score performance,

We acknowledge that the performances of the binary and full data optimal scores relative to that of the sum score is likely to vary with test length, the distribution of the test scores across examinees, the nature of the knowledge or other state being tested and so forth. Our future work will involve constructed answer tests and rating scales as multiple choice tests, and will reveal how these other test- and examinee-specific features affect score performance.

We have developed a computer application called *TestGardener* that requires almost no knowledge of item response theory or statistics to use (Li, J., Ramsay, J. O. and Wiberg, M., 2019) The application is available from the authors, and can be also accessed through our current web server <http://testgardener2019.asurewebsites.net/>.

References

- Lee, Y-S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement*, 37, 121-134.
- Li, J., Ramsay, J. O. and Wiberg, M. (2019) TestGardener: A Program for Optimal Scoring and Graphical Analysis. In Wiberg, M., Culpepper, S. A., Jansen, R., Gonzlez, J., and Molenaar, D. (Eds.) Quantitative Psychology 83nd Annual Meeting of the psychometric society, New York, NY, 2018, New York: Springer. 87-94.
- Ramsay, J. O. (1997). A functional approach to modeling test data. In W. J. van der Linden and R.K. Hambleton (Eds.), *Handbook of modern item response theory* (351-367). New York: Springer.
- Ramsay, J. O. Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. New York: Springer-Verlag.
- Ramsay, J. O. and Silverman, B. W. (2002). Functional models for test items. In *Applied functional data analysis*, Chapter 9, (pp. 131-144), New York: Springer-Verlag.
- Ramsay, J. O and Wiberg, M. (2017). A strategy to replace sum scoring. *Journal of Educational and Behavioral Statistics*, 42(3), 282-307.
- Rossi, N., Wang, X. and Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of the Behavioral and Educational Sciences*, 27, 291-317.

- Tribus, M. (1961) *Thermodynamics and Thermostatistics: An Introduction to Energy, Information and States of Matter, with Engineering Applications*. New York: Van Nostrand.
- Wiberg, M., Ramsay, J. O. and Li, J. (2018a). Optimal scores as an alternative to sum scores. In Wiberg, M., Culpepper, S. A., Jansen, R., Gonzalez, J., and Molenaar, D. (Eds.) *Quantitative Psychology 82nd Annual Meeting of the psychometric society*, Zurich, Switzerland, 2017, New York: Springer. 1-10.
- Wiberg, M., Ramsay, J. O., and Li, J. (2018b). Optimal scores: an alternative to parametric item response theory and sum scores. *Psychometrika*.
<https://doi.org/10.1007/>
- Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, 11, 253-270.
- Woods, C. M., and Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities, *Psychometrika*, 71(2), 281-301.