Optimal scores: an alternative to parametric item response theory and sum scores

Marie Wiberg, James O Ramsay, & Juan Li

Acknowledgements

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

Published paper. Please cite as:

Wiberg, M., Ramsay, J. O., & Li, J. (2019). Optimal scores: an alternative to parametric item response theory and sum scores. *Psychometrika*, *84*(1), 310-322
https://doi.org/10.1007/s11336-018-9639-4

Optimal scores: an alternative to parametric item response theory and sum scores

Abstract

The aim of this paper is to discuss nonparametric item response theory scores in terms of optimal scores as an alternative to parametric item response theory scores and sum scores. Optimal scores take advantage of the interaction between performance and item impact that is evident in most testing data. The theoretical arguments in favor of optimal scoring are supplemented with the results from simulation experiments, and the analysis of test data suggests that sum-scored tests would need to be longer than an optimally scored test in order to attain the same level of accuracy. Because optimal scoring is built on a nonparametric procedure, it also offers a flexible alternative for estimating item characteristic curves that can fit items that do not show good fit to item response theory models.

1. Introduction

Test scores are used to make decisions about test takers, and thus it is important that such scores estimate the ability levels of the test takers accurately. In the past, the use of sum scores (i.e. the number correct) has been the primary method of scoring tests, although parametric item response theory (IRT; Lord, 1980) is used, but primarily for representing item characteristics. Sum scores have the advantages of being easily calculated, computationally fast, and easy for the test takers to understand. A problem with using parametric IRT is that not all items can be satisfactorily modeled with a parametric IRT model, even in well-designed tests. The aim of this paper is to discuss the nonparametric IRT approach of optimal scoring and to compare it with parametric IRT and sum scoring using data from the Swedish Scholastic Assessment Test (SweSAT).

In the usual test theory, ability θ is represented as a signed real number, and the one- (1PL), two- (2PL), and three-parameter logistic (3PL) IRT models are used to model the probability $P_i(\theta)$ that item i will be answered correctly at that ability level (Birnbaum, 1968). The logistic family of item characteristic curves (ICC) are defined for the 3PL as

$$P_i(\theta) = c_i + \frac{1-c_i}{1+e^{-a_i(\theta-b_i)}}, \qquad (1)$$

where a_i is the item discrimination, b_i is the item difficulty and c_i is the pseudogessing parameter.

However, students and their teachers are apt to prefer the familiar score intervals $[0,n]$ or $[0,100]$ for representations of either ability or actual performance, and would find an ability estimate of, say, -0.15 difficult to comprehend. Since θ is a latent variable and thus not actually observed, then for any ψ in, say, $[0,100]$ there exists a function $P_i^*(\psi)$ such that the two probability values are equal. This can be achieved, for example, by the scaled logistic transformation

$$\psi = \frac{100e^{\theta}}{1+e^{\theta}}, \tag{2}$$

and an infinity of other transformations exist as well. Of course the formulas for these three models will change as a consequence, but as long as the transformation preserves rank order and is reasonably smooth, this will be an acceptable score metric. Note, however, that using the whole real line will make transforming abilities of $-\infty$ and ∞, which are logical expressions of zero or perfect test scores, respectively, awkward from a computational point of view.

An alternative to parametric models is to use nonparametric methods to estimate abilities and ICCs, and there have been many proposals for such ICCs in the past. Mokken (1997) studied nonparametric estimation in connection with monotonicity. Ramsay (1991, 1997) proposed an ICC estimation with kernel smoothing over quantiles of the Gaussian distribution, which gave fast and reasonably accurate ICC estimations in the program TestGraf. Ramsay and Silverman (2002) and Rossi, Wang, and Ramsay (2002) used the expectation-maximization (EM) algorithm to optimize the penalized marginal likelihood, and the ICC estimates came close to the 3PL IRT model when their roughness penalty increased. Woods (2006) and Woods and Thissen (2006) simultaneously estimated item parameters using a spline-based approximation of the ability distribution. Ramsay and Silverman (2005) proposed a method for nonparametric, but not strictly monotonic, curve estimates. Later, Lee (2007) compared several nonparametric approaches with each other.

We define an estimated score as *optimal* if it optimizes some criterion for fitting the data provided by a test taker within some class of IRT models. In this paper, we have chosen to use maximum likelihood estimation (MLE) of a model, which is capable to represent the data as accurately as desired. To use MLE is in general considerable to be asymptotically efficient.

If the criterion is negative log likelihood and the data are binary scores $U_i, i = 1,...,n,$ for example, then the fitting criterion for an arbitrary item response function over an arbitrary score interval is

$$-\log L(\theta \mid U) = \sum_{i=1}^{n}[-U_i W_i(\theta) + \log(1 + e^{W_i(\theta)})] \tag{3}$$

where the log-odds function $W(\theta)$ is defined as

$$W(\theta) = \log\frac{P(\theta)}{1 - P(\theta)}.$$

In this paper, we focus on optimal scores over intervals [0,n] or [0,100] using the nonparametric item response functions in Ramsay and Wiberg (2017a) computed using spline-based expansions of item response functions. The use of these nonparametric expansions enables fast computation of item response function estimates, which can fit the data as closely as desired. This paper differs from Ramsay and Wiberg (2017a) in that it provides a nose-to-nose comparisons between nonparametric sum scores and optimal scores estimated using either parametric or nonparametric ICCs.

The rest of this paper is structured as follows. In the next section we introduce the SweSAT data and examine the empirical distribution of the data and the sum scores. In the third section, we move to the estimation of ICCs and the different test scores, including a brief description of the estimation of nonparametric ICCs. In the fourth section, optimal scoring is compared with sum scoring and parametric IRT scores. The paper ends with a discussion and some concluding remarks.

2. Sum scores as a point of departure

The SweSAT is a college admissions test with multiple-choice items that are binary scored. The SweSAT consists of a quantitative and a verbal part, both having 8o items. In this paper, only the quantitative part was used. Sum scores are typically used, and these are then transformed into scaled scores that are comparable to other administrations of the test. We will denote sum scores with S_j, $j = 1,...,N$. In the left panel of Figure 1, the empirical distribution of the sum scores and the later-described optimal scores for 30,000 test takers who took the SweSAT are shown. It is noticeable that neither scores are normally distributed.

The right panel in Figure 1 contains estimates of the 80 item response functions for the SweSAT estimated by the Ramsay and Wiberg (2017a) nonparametric procedure. It is obvious that the SweSAT is difficult for the majority of the test takers; the median score was 35, the lowest score was 4, and only a single individual achieved a perfect score. From the right panel of Figure 1, it is evident that some items have rather flat ICCs and that some items are much more difficult than others. We especially note that a number of ICCs appear to reach flat upper asymptotes that are well below the value one predicted from the 3PL model, suggesting that these items, failed by even the brightest test takers, are in some way defective. If so, the sum score is then also defective since there is no assurance that a near perfect score is achievable. This highlights the main limitation of the sum score, which is that it takes no account of item performance, or of any interaction between item and test taker performance. That is, a low sum score may reflect poor item performance as well as limited ability, in which case the test taker is blamed for what is essentially not his fault by the use of the sum score.

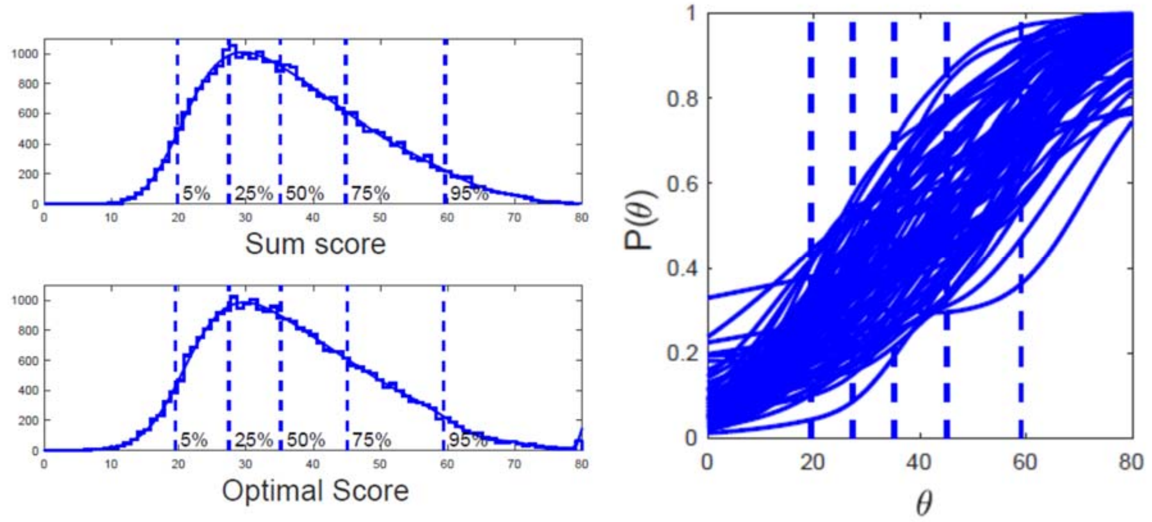

Figure 1. The left panel shows the empirical distributions of sum scores and optimal scores for the SweSAT and a smooth density function rescaled over [0,80] to overlay them. The right panel shows the $P_i(\theta)$ curves estimated over the interval $\left[0,80\right]$. The vertical dashed lines are the 5%, 25%, 50%, 75% and 95% quintiles of the empirical distribution of the sum scores.

3. Estimating the log-odds functions $W_i(\theta)$

Ramsay and Wiberg (2017a) first proposed nonparametric IRT scores in terms of optimal scores. Instead of estimating $P_i(\theta)$, they concentrated on estimating the log-odds function $W_i(\theta)$. Because θ is defined over a closed interval and the values of $W_i(\theta)$ are unbounded, an efficient estimate of $W_i(\theta)$ uses B-spline basis function expansions

$$W_i(\theta) = \sum\nolimits_k^K \gamma_{ik}\phi_{ik}(\theta), \qquad (4)$$

where $\phi_k(\theta) = B_k(\theta \mid \xi, d)$, γ_{ik} is the coefficient of the basis function, $B_k(\theta \mid \xi, d)$ is the B-spline basis function, ξ is a knot sequence, *k* is the number of spline functions, and *d* is the order of the spline.

Figure 2 displays in its upper left panel $W_i(\theta)$ for a 2PL ICC (*a=1, b=c=0*) and in the lower left for a 3PL ICC (*a=1, b=0 ,c=0.25*). We see that the log-odds transformation tends to convert probability curves into straight or only mildly curved shapes, which renders the task of approximating them much easier. However, when we transform θ into percentage values using the scaled log-odds transformation, we see that both extremes for the 2PL model approach vertical asymptotes and the right extreme does so for the 3PL model. The left panel of Figure 3 shows the $W_i(\theta)$ curves for the SweSAT data, which correspond to the $P_i(\theta)$ curves in the right panel of Figure 1. There we do not see any indication of asymptotic behavior for either extreme, and instead only a mild tendency to curvature. The *W* curves are more informative than the *P* curves because the two horizontal asymptotes of probability hide important information about how rapidly these asymptotes are being approached.

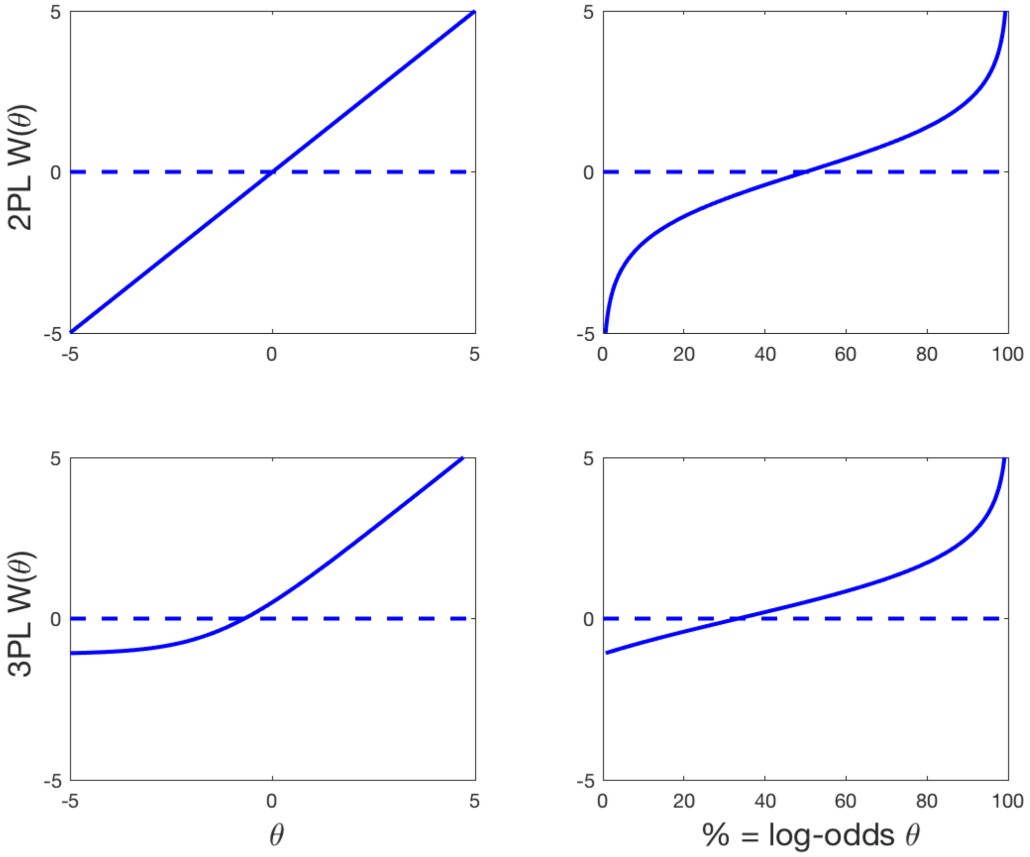

Figure 2 The upper and lower left panels show the log-odds transformations of the item characteristic curves for 2PL and 3PL models, respectively. The right panels display these

curves when the argument θ is mapped into the interval [0,100] using the scaled version of the log-odds transformation.

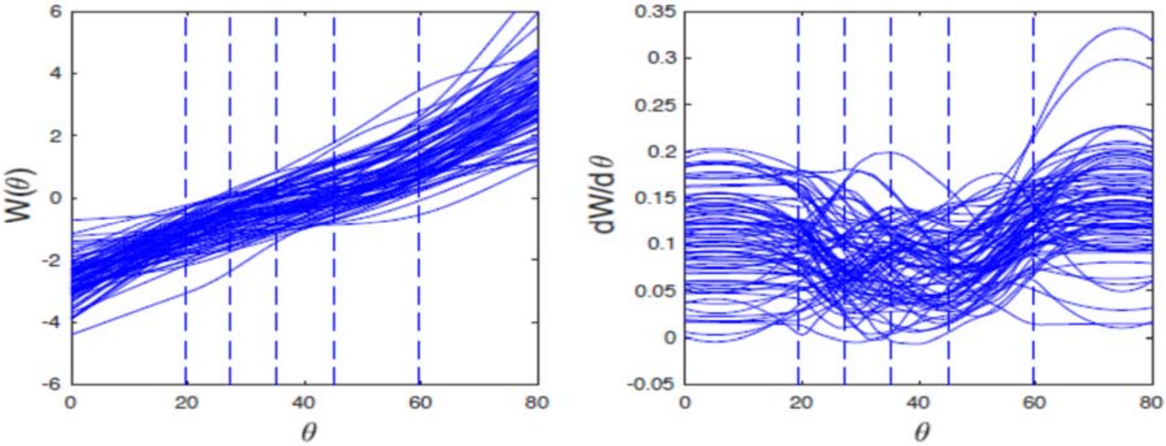

Figure 3. The estimated log-odds functions $W_i(\theta)$ for the SweSAT are displayed in the left panel. The item impact curves, $dW_i / d\theta$, which provide the optimal weighting of the item scores, are displayed in the right panel.

The log-odds function $W(\theta)$ is also preferable to the probability-valued ICC because of its pivotal role in the negative log likelihood equation (3), and we see its importance even more clearly in the first derivative of the negative log likelihood defined at the optimal θ

$$\sum_i^n U_{ij} \frac{\mathrm{d}W_i}{\mathrm{d}\theta} - \sum_i^n P_i(\theta) \frac{\mathrm{d}W_i}{\mathrm{d}\theta} = 0 . \tag{5}$$

The first of the two terms on the left side is a weighted sum of the data, the second is the same type of sum of the fit, and at the optimal θ the two terms are equal. The *n* weights $dW_i / d\theta$ are the slopes of the log-odds functions at θ, and determine the importance of each term in the sums. That is, the quality of the information provided by a response is measured by how fast the log-odds is increasing at θ; so that where $W_i(\theta)$ is flat the response is completely uninformative. This tends to happen when the test taker either finds the item too easy to challenging or too difficult to permit any strategy other than guessing. On the other hand, locations where $W_i(\theta)$ is sharply increasing are those where the response matters a great deal in estimating the test taker's ability.

If we know which wrong option a test taker chooses, this will provide additional information about θ because some wrong options are more wrong than others. If data are available for which M options are given, the scoring accuracy can be improved by modeling the option choices using the functions $W_{im}(\theta),\ m=1,...,M_i$. Thus, the multinomial version of Eq. (5), where we are not bundling the wrong options into a single failure category, becomes

$$\sum_i^n\left[\sum_m^{M_i} U_{jim}\frac{dW_{im}}{d\theta_j}\right]-\sum_i^n\left[\sum_m^{M_i} P_{im}\frac{dW_{im}}{d\theta_j}\right]=0\,, \qquad (6)$$

where $dW_{iU_{ji}}/d\theta$ is the slope of the option for item *i* chosen by test taker *j* at ability θ. Here, too, we see a weighted sum over residuals, where the second term is the expected *W(θ)* slope that is required on the average to fit the data in the first term. The central idea captured by both Eq. (5) and Eq. (6) is that items vary in the shape of their log-odds functions $W_i(\theta)$, as is obvious from the left part of Figure 3.

The 80 item slopes of the log-odds curves are shown for the SweSAT data in Figure 3. From the right panel we see that there are sub-groups of items that switch from high to low impact and from low to high impact. In other words, there is an interaction between performance and item impact that is missing in an a priori weight assignment, such as the unit weight that defines the sum score. The term “performance-sensitive scoring” might be a useful description of optimal scoring, and we might refer to the slope-log-odds functions as *item impact curves*.

An interpretation of optimal scoring motivated by linear regression is that Eq. (5) is a continuum of regression models indexed by θ for which the residuals $U_{ij}-P_i(\theta_j)$ are predicted by the single covariate $dW_i/d\theta_j$, and the value of θ_j is chosen to satisfy the requirement in regression analysis that the residuals be orthogonal to the covariates. In the multi-option version of this equation, too, we see a weighted sum over residuals, where the second term is the expected *W* slope that is required on the average to fit the data in the first term.

Ramsay and Wiberg (2017a) point out that a simple adequately accurate method for estimating the log-odds functions is to bin the score values S_j into K bins such that the numbers of scores in the bins are roughly equal, and then use spline smoothing over bin centers of the log-odds transformations of the bin proportions (Ramsay, Hooker and Graves, 2009). This took under a second for the SweSAT data.

The distribution of *θ* is determined by the shapes of the estimated probability or log-odds functions, and since the initial smoothing uses sum scores as the abscissa values, the estimated ability distribution will resemble that of the sum scores as shown in Figure 1. This operation may be repeated but using optimal score estimates rather than sum scores in order to define a distribution that more closely resembles that for the optimal scores.

4. Comparing different test scores

4.1 Optimal scoring in comparison with sum scoring

If we wish to consider moving from sum scores to optimal scores, we should examine to what extent optimal scores deviate from sum scores. Figure 4 shows a contour plot of the density $\theta_j - S_j$ as a function of S_j for the SweSAT data. The differences might be as much as 20% of the score values near the median, but particularly striking is the tail to the right in the plot showing a considerable increase in scores for the high-performance test takers. In other words, high achievers would obtain a higher score if optimal scoring were used instead of sum scoring.

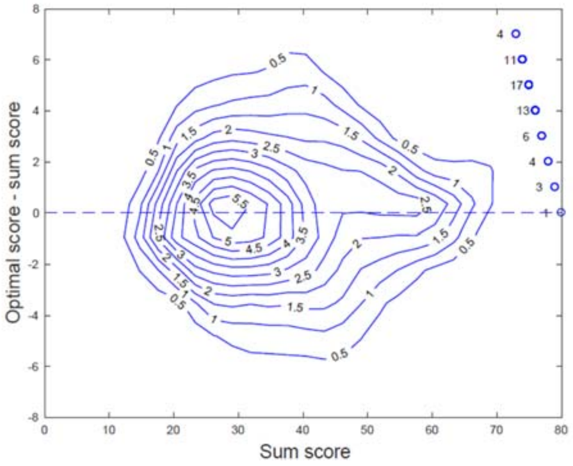

Figure 4. A contour plot of the density of the optimal score minus the sum score as a function of the sum scores for the SweSAT data. The points lined up on the left of the contours are test takers who were assigned scores of 80, and the numbers indicate the number assigned.

4.2 Optimal-scoring ICC in comparison with parametric IRT ICC

Parametric IRT is widely used among test constructors because it is believed that, given one of these parametric families, the estimated scores are invariant over different samples. Although many items can be modeled with a parametric IRT model, a problem is that even in a well-designed large-scale assessment such as SweSAT there are some items that do not fit a parametric IRT model. The different lines in the two SweSAT items in Figure 5 correspond to an ICC from a 3PL IRT model and from spline smoothing. To generate these figures, we first binned the observed data in 25 bins containing equal numbers of test takers. Thus, the circles in the figures represent the middle values of each of the bin intervals. Clearly, nonparametric IRT in terms of optimal scoring allows us to model items that do not have a good fit to commonly used parametric IRT models. Of course, it is appropriate to suggest that all we have to do is to add a parameter or two to provide more flexibility, but the well-known difficulty of computing the lower asymptote parameter is likely to persist, and the spline curves shown in the figure involved only five parameters and their estimation of these parameters is a far faster and more stable process.

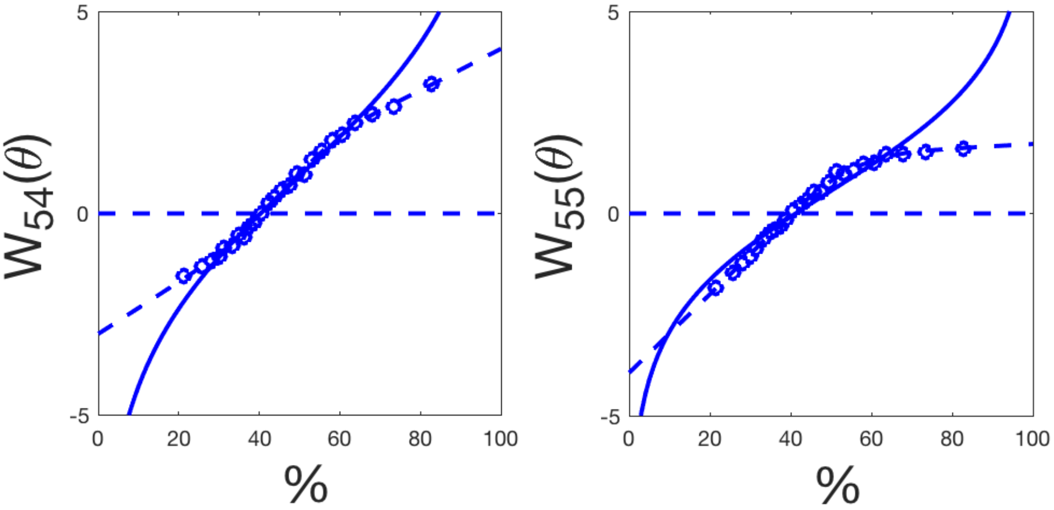

Figure 5. The fits to the log-odds data for two SweSAT test items (items 54 and 55) using the 3PL model (solid line) and spline smoothing (dashed line).

4.3 Comparison of the three different test scores

To explore how the different test scoring methods perform, we compared the efficiencies of sum scores, parametric IRT scores, and optimal scores. We simulated data where the population was defined by the W_i curves and the θ_j s estimated from the SweSAT data. To assess recovery of θ, the root mean squared error (RMSE) of θ was used because the RMSE is composed of both the bias and sampling variance. We also examined the average bias over θ. We did not use correlations because correlations are a global measure, and work best if we have a linear relationship, which was not the case here. There are two estimation situations: (1) the functions $W_i(\theta)$ have been pre-calibrated and can be taken as known, or (2) the same data are used to assess test takers and to estimate item characteristics. In our simulation analyses, we found that the two types of analysis are practically indistinguishable for $N > 1000$. The RMSE of θ inevitably increases if it has to share information with the parameters that define the W_i functions, but our experience suggests that the RMSE can still be a useful statistic for N on the order 400 or so (Ramsay & Wiberg, 2017a).

Our simulation strategy had to deal with the identification of the distribution of θ. In order to give sum scores the maximum advantage, we simulated test data using a smooth estimate of the density of the sum scores based on an actual set of data as the template distribution. Because we had 30,000 test takers, we could assume that the W_i functions were known (i.e. situation 1 above) and that we only needed to simulate the test takers' responses. The analyses of both the real and simulated data used the estimation of log-odds function approach given in Ramsay and Wiberg (2017a). The results reported here are based on simulations that emulated the SweSAT, and consequently Figures 1 and 3 can be consulted for information about the test characteristics. In particular, it should be noted that a number of items failed to reach $P_i = 1$, thus sum scores near 80 are rare.

We simulated 1,000 test takers' responses and used the 81 sum score values as fixed values of θ so that we could examine how RMSE varied over θ. For each value of θ, sum scores, optimal scores computed from spline estimates of log-odds functions, and optimal scores computed from 3PL IRT scores were averaged across 1,000 simulated samples. The Matlab code used in the simulation study is given as a supplementary file. The top and bottom panels of Figure 6 show the average bias and the average RMSE for the three estimates as a function of θ, respectively. For the central 90% of the test takers, the optimal scores and sum scores are essentially unbiassed, the 3PL scores exhibit some bias, but the sum score and 3PL biases are much larger for the 10% of the most extreme scores. Low-end test takers get a boost from the sum scores and optimal 3PL scores, but the high-performing test takers lose as much as five items relative to the θ values used to generate the data.

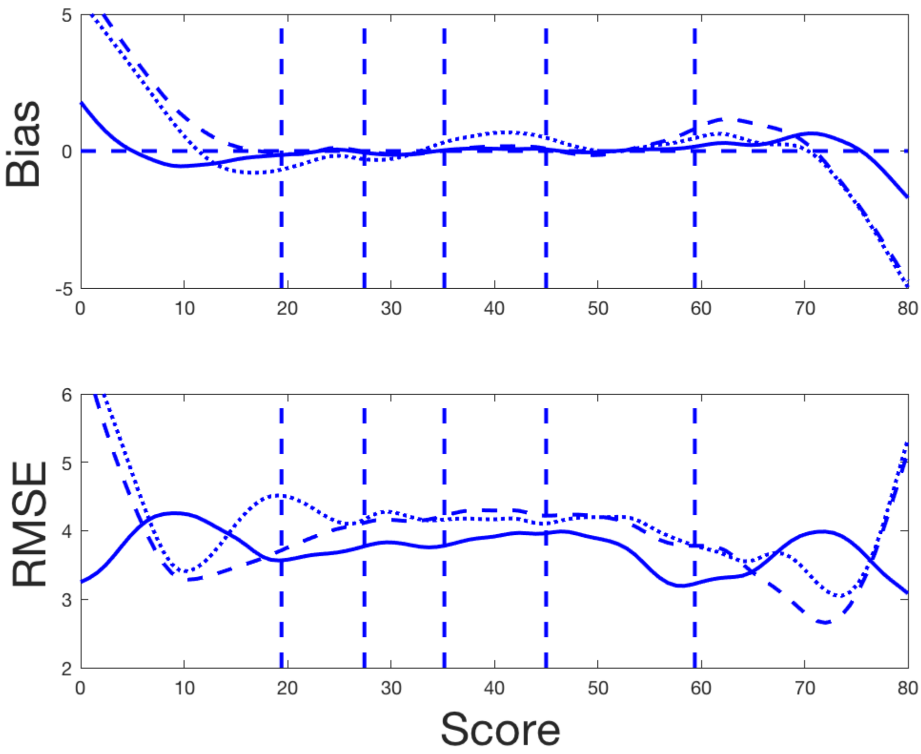

Figure 6. The top panel shows the average bias of θ estimates, and the bottom panel shows the average RMSE of θ estimates for the SweSAT data. The solid lines are for optimal scores estimated from nonparametric log-odds estimates, the dashed lines are for sum score estimates and the dotted lines are for 3PL log-odds estimates.

The RMSE's for the sum and optimal 3PL scores are larger than the RMSE's for optimal nonparametric scores for the majority of the central 90% of test takers. The average improvement of the optimal nonparametric score RMSE over the sum score RMSE is 6.8%, corresponding to a mean squared error (MSE) improvement of 14%. Because the MSE declines in proportion to $1/n$, we see that the sum-scored SweSAT would have to be 14 items longer than an optimally scored test in order to achieve the same average accuracy. For test takers not too far from the 5% quantiles, the weighting used for optimal scoring effectively decreases the amount of information available for estimation and therefore has the larger RMSE. On the other hand, for test takers with nearly zero or perfect scores, the biases dominate the sampling variance and inflate the RMSE. The tail on the right of the contour plot in Figure 4 shows how optimal scoring compensates high-performance test takers for the negative bias in sum scores. In particular, 49 test takers with sum scores ranging from 73 to 80 were assigned optimal scores of 80. We believe this to be realistic given the size of the sample. These results for optimal scores compared with sum scores are in line with those in Ramsay and Wiberg (2017a) who showed simulation results over the score intervals for three exams using the scenario where there is joint estimation of W_i functions and optimal scores.

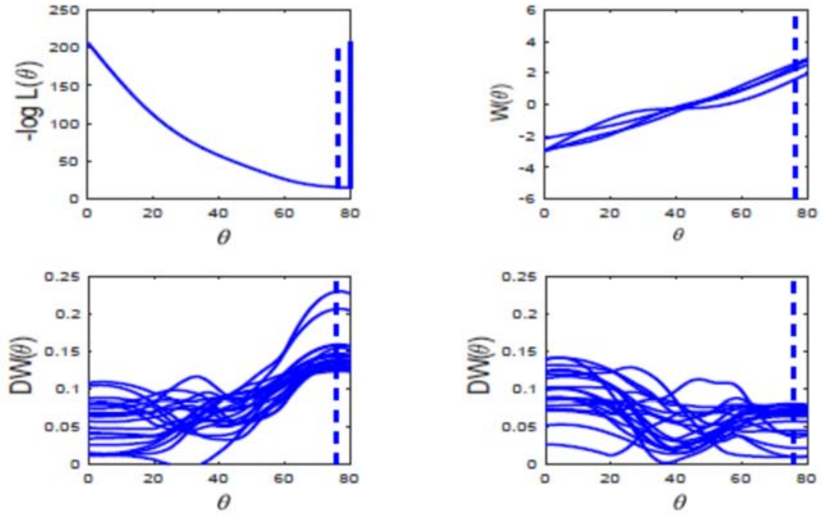

Figure 7. Results for a test taker with a sum score of 76 and an optimal score of 80. The upper left panel shows the negative log likelihood function, which is minimized at a score of 80. The dashed line shows the sum score for this test taker, and the solid line shows the optimal score. The upper right panel shows the W_i functions for the four items that this test taker answered incorrectly. The lower left panel shows the 20 W_i derivatives with the highest values at a score of 76, and the lower right panel shows the derivatives for the 20 W_i derivatives with the lowest values.

The task of explaining the differences between optimal scores, parametric IRT scores, and sum scores to test takers who are accustomed to sum scores will be challenging. Figure 7 shows what happens for a test taker who answered four items incorrectly but nevertheless was assigned an optimal score of 80, as shown in the upper right panel. The W_i functions for the incorrect items in the upper right panel are not particularly informative, and at a score of 76 these are still well below the best curves in Figures 1 and 2. Thus it is not unusual for a high-performance test taker to get items like these wrong. The bottom two panels show the W_i derivatives for curves with the highest and lowest values at a score of 76. In particular, we see in the lower left panel in Figure 7 that these high weights are more or less constant between scores of 76 and 80, and consequently the maximum likelihood estimate criterion is tending to treat all test takers in this elite range the same.

5. Discussion

The overall aim of this work was to show how nonparametric IRT in terms of optimal scoring is preferable to sum scoring and even optimal parametric IRT scoring. To do this, we used a sample from one administration of the SweSAT. From Figure 1 we can see that the empirical distribution is somewhat right skewed. It has been shown previously that the distribution of θ is arbitrary, but this can be turned into an asset because it allows us to model ICCs over any interval. Here we chose to score them over closed intervals because the results are easier to interpret than if the whole real line is used. By defining θ over the range of sum scores or percentage, we can make a nose-to-nose comparison of their efficiencies and can communicate the results in more understandable terms to the test takers.

In the sample data used here, the average RMSE and average bias were lower for optimal scores and thus such scores are to be preferred everywhere over the sum scores. A problem with the sum score is that it is severely positively biased for the weakest test takers and negatively biased for the strongest test takers. Bias in statistics is usually considered intolerable except in situations where it can greatly decrease RMSE. In an admissions test like SweSAT this is especially harmful as top achievers might still not get accepted to the university program of their choice. We can also expect greater benefits from using optimal scoring for tests developed in messier environments, for example, intermediate-sized government agencies, such as in the case of the four provincial exams given to all students in the province of Ontario, Canada.

An attractive feature of optimal scoring in comparison to parametric IRT scoring is the added flexibility of using smoothing estimates of the log-odds functions $W_i(\theta)$. Low-dimensional models like the 2PL or the 3PL IRT models might look nice, but the data do not always agree with such models, especially if the underlying distribution differs from the assumptions the models rely on. The sum scores are still useful in some situations, but if we put some effort into explaining how the maximum likelihood-weighted score works, optimal scores should be seen by test developer and test taker as preferable. We are, however, well aware that switching from sum scores to optimal scores will not be easy in practice. Parametric IRT is useful if we have test data that agree with the chosen model, but our practical experience as well as the SweSAT example suggest that there are always items – even in well-designed tests – that do not fit a parametric IRT model very well.

The MSE and RMSE require that the point estimates are on the same scale, thus a point-to-point comparison of the sum score to a competitor score requires that the two scores be on the same metric. The recasting of psychometric theory into closed intervals anchored at 0 seems essential to make comparisons and provides a natural medium for discussing performance estimates with test takers and test constructors. Our results show that even well-designed tests can be improved in terms of accuracy if optimal scoring is used, and the average improvement in test scores is about 6% for 90% of the test takers. An improvement in the test score might matter a lot for the best test takers, and because the SweSAT is a college admissions test this could be the difference between being accepted to the university program of one's choice or not.

Our improvement in assessment accuracy rests on the combinations on four features, each of which has technical as well cosmetic implications. First, transposing to the domains [0,n] and [0,100] that are familiar to everyone permits us to use B-spline bases to represent both the option curves and the density of the scores. It also enables a point-wise assessment of bias and MSE for the optimal and sum scores rather than the often-used correlation criterion, which is only a global measure of linear agreement. This in turn highlights the bias issues for extreme scores and the sampling variance issue for moderate scores. Secondly, we use the log-odds or logit transform to map observed proportions into values $W(\theta)$ with an unrestricted range. Thirdly, by binning the ordered scores so that they have roughly equal frequencies, we can use an extremely fast spline data-smoothing algorithm that can fit the log-odds data to an arbitrarily high level of precision and which is relatively stable in the low data-density extremes of the ability continuum. We can thus fit well even ill-behaved items which are usually not fit very well by parameter IRT models. Finally, the derivative of this log-odds transformation is a point-wise measure of the discriminability of the item at a specific score value that defines the weights in a differentially weighted sum score that in turn defines the maximum likelihood estimate of an examinee's ability conditional on the estimated logit curves.

We refer to this package as optimal score estimation, not in the sense that it is the best possible, but because it optimizes the likelihood of an examinee's response sequence using an item response model that is as accurate as we choose. By contrast, the sum score does not optimize any criterion, including the likelihood of the Rasch model, even in the unlikely circumstance that it could provide a reasonable fit to the data. Of course, the maximum likelihood criterion assumes local independence and that the true score lies within a compact set. In our on-going development, we expect to be able support deviations from these assumptions as well.

Although we show in this paper that the use closed-interval analogues of popular parametric models defined over [-∞,∞] do not have the capacity to fit that data that we analyze, we are left with the interesting question of how many parameters a parametric model would need to have to provide a comparable fit to the data. Our use of functional principal components analysis of the log-odds curves suggests that about six would be required, and indeed, we see that using six B-spline basis functions also does well in representing the log-odds curves and supporting optimal scoring.

Future research can go in many different directions. Where deemed appropriate, covariates are often available that shed further light on the state of the test taker. The use of covariates such as age, gender, educational background, and so on is not encouraged in academic testing, although there are recent examples of using covariates successfully in test equating (Wiberg and Bränberg, 2015). Covariates are certainly important in clinical assessment contexts provided that disclosure issues and other considerations are properly dealt with. Future research should also extend the proposed approach to polytomously-scored items and to multidimensional tests. It is also important to examine different types of tests. In this paper we had almost no one scoring at or near the endpoints and it is thus of interest to examine a diversity of tests and compare the results to other scores.

Another important step is to make the results available to other researchers and test constructors. We are currently developing a new version of the TestGraf program described by Ramsay (2000) for both Microsoft and Apple operating systems that aims to provide much of this capability. We also envisage software that is executable on mobile devices and available through bubble-sheet scoring services. The test takers, who spend hours providing the data, have a right to know not only their final score, but also how it was obtained, what alternative scoring methods were available, confidence limits, score distributions, and other inferential material to help them understand their performance on the test. For example, full disclosure of the SweSAT items is required in Sweden. The voices of parents of gifted children who realize what the tail in the contour plot in Figure 4 implies are apt to be a powerful aid in selling the idea of performance-sensitive testing. The simulation analyses presented here, especially in comparison to sum scores, should help to persuade test administrators to use optimal scoring of their data. A future challenge is however to describe for test taker what optimal scores are and to make them understand how they work.

An important step in making optimal scoring more mainstream is to construct simulation applications that can accept real data in a wide range of formats as discussed in Ramsay and Wiberg (2017b). Those can minimize the work involved in setting up a simulation based on these real data, that are fast enough to permit demonstrations within a meeting situation, and that come equipped with the kind of graphical display and automatically generated documentation that executives and managers will find useful. The principles for such software have already been demonstrated by applications in both the R and Matlab communities, but our own progress in this direction is limited. It should, however, be doable because the proposed approach has proven to be computationally fast. The simulation experiment with the SweSAT data only took about four minutes with an iMac with four processors running in parallel.

We began our research with the question, “Can we improve the scores given to test takers?” We are immensely excited by the opportunities that come with optimal scoring, and we believe that further use of such methods will show that reducing testing data to a binary form wastes a significant amount of useful information. Moreover, we believe that, as valuable as the unidimensional model for multiple-choice tests has been, it has its limitations and it is worth taking the leap towards using new and more efficient statistical methodologies.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F.M. Lord & M. R. Novick (Eds.). *Statistical theories of mental tests* (pp. 395-479). Reading, MA: Addison-Wesley.

Lee, Y.-S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 37*, 121–134.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp.351–367). New York, NY: Springer.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630.

Ramsay, J. O. (1997). A functional approach to modeling test data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York, NY: Springer.

Ramsay, J.O. (2000). *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data.* [Computer software and manual]. Department of Psychology, McGill University, Montreal Canada.

Ramsay, J. O., Hooker, G. and Graves, S. (2009). *Functional data analysis in R and Matlab.* New York, NY: Springer.

Ramsay, J. O., & Silverman, B. W. (2002). Functional models for test items. In *Applied functional data analysis.* New York, NY: Springer-Verlag.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis*. New York: Springer.

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 scoring barrier*. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W-C. Wang. (Eds.) Quantitative Psychology – 81st Annual Meeting of the psychometric society, Asheville, North Carolina, 2016, New York: Springer. 151-158.

Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of the Behavioral and Educational Sciences*, 27, 291–317.

Wiberg, M. & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement,* 39(5), 349-361

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., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71*, 281–301.