<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 O., Wiberg, M. (2017)

A Strategy for Replacing Sum Scoring

Journal of educational and behavioral statistics, 42(3): 282-307

<https://doi.org/10.3102/1076998616680841>

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-136047>

A Strategy for Replacing Sum Scoring

James Ramsay (corresponding author)
Department of Psychology, McGill University
2748 Howe St., Ottawa, ON, K2B 6W9
Canada
James.Ramsay@mcgill.ca

Marie Wiberg
Department of Statistics
UmeåUniversity
UmeåSE-901 87
Sweden
marie.wiberg@umu.se

James Ramsay is a Professor Emeritus in the Department of Psychology and an Associate Member of the Department of Mathematics and Statistics at McGill University, 2748 Howe St., Ottawa, ON, K2B 6W9, Canada, email James.Ramsay@mcgill.ca. His research interests are psychometrics, functional data analysis, dynamic systems identification and spatial data analysis.

Marie Wiberg is a Professor at the Department of Statistics, Umeåschool of Business and Economics, UmeåUniversity, SE-901 87, Sweden, marie.wiberg@umu.se. Her research interests include psychometric models and methods, test equating and international large-scale assessments.

This paper has been published. Please cite as: Ramsay, J. O. & Wiberg, M. (2017). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics*, 42(3), 282-307. <https://doi.org/10.3102/1076998616680841>

A Strategy for Replacing Sum Scoring

J. O. Ramsay

Department of Psychology, McGill University

M. Wiberg

Department of Statistics, Umeå University

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.

Abstract

This paper promotes the use of modern test theory in testing situations where sum scores for binary responses are now used. It directly compares the efficiencies and biases of classical and modern test analyses, and finds that an improvement in the root mean squared error of ability estimates of about 5% for two designed multiple choice tests and about 12 % for a classroom test. A new parametric density function for ability estimates, the tilted-scaled beta, is used to resolve the non-identifiability of the univariate test theory model. Item characteristic curves (ICC's) are represented as basis function expansions of their log-odds transforms. A parameter cascading method along with roughness penalties is used to estimate the corresponding log-odds of the ICCs, and is demonstrated to be sufficiently computationally efficient that it can support the analysis of large data sets.

Key Words: parameter cascading, item characteristic curves, tilted scaled beta distribution, sum score distribution, performance manifold

1 Introduction

1.1 Preliminaries and overview

Psychometricians classify statistical methodology for the analysis of testing data as classical or modern. The primary difference between the two is simple: the classical approach, well presented in Gulliksen (1950), presumes a fixed scoring scheme and proposes a variance components model for the scores. Most performance tests are analyzed by summing fixed a priori item scores, which we refer to as the *sum score*, and for multiple choice tests the sum score is the number of correctly answered items.

The modern theory, on the other hand, works with a weighting of item scores where the weights vary over the estimated abilities. It is clear from even a casual inspection of testing data that test items vary in quality. For example plotting the proportion P_i of correct answers for a single item i against the sum scores $S_j, j = 1, \dots, N$ of examinees is an effective strategy for detecting items for which P_i scarcely varies in level for the great majority of examinees or, all too often, even decreases. The essence of modern testing data analysis is the modelling of such a plot by the relation $P_i(\theta)$ where θ is a measure of performance.

The principal contribution of item response theory (IRT) to testing has been the use of the item characteristic curve (ICC) as an indicator of item performance in the test design process by large commercial and government testing agencies. However, the use of IRT to improve test scoring has had much less impact, and especially in more informal testing situations as school and university classrooms. Even industrial level testing agencies such as Educational Testing Service use the sum score, and justify the practice by carefully designing and screening their items are so as to make the unweighted sum score as close to optimal as possible. Within academic institutions, where tests are designed by teachers and professors, multiple choice item scores, as far as we are aware, are virtually always summed; and, although some tests and scales do have weighted items, these weights are assigned a priori by the tester rather than estimated in an optimal manner.

Figure 1 displays for illustration purposes the estimated ICC's $P_i(\theta), i = 1, \dots, n$, for two designed tests: (1) the norm-referenced Swedish Scholastic Aptitude Test (SweSAT), an admission test for post-secondary education in Sweden with 122 items administered to a single sample of 2766 examinees; and (2) the American history portion of the National Assessment of Educational progress (NAEP) with 36 items administered to 2440 examinees. While most ICCs have the shape characteristics

of the three-parameter logistic curves often used in statistical test theory analyses, some exhibit right asymptotes that are well below unity and multiple inflection points. More details on the analyses that yielded these curves are found in Section 5.

Insert Figure 1 here.

Figure 2, unfortunately, displays a far more typical situation. The 100-item multiple test was designed by a professor with an outstanding international reputation for his research, and a passion for excellent teaching that has attracted generations of first year students to major concentrations in his discipline. One item, 96, discriminates *against* the best students in the class, and five others are flat-lined at values that also penalize the best students. It is therefore not surprising that the highest sum score was only 91 and that only 5% of the students scored above 83. Many other items are answered correctly by nearly everyone, and therefore contribute information that is both uninformative and positively biased. To make matters worse, after some decades of assigning students in a course on psychometrics to prepare reports on instructor-prepared examinations, one of the authors can assert that this fraction of poor items is relatively small, so that the test is well above average in quality.

Insert Figure 2 here.

So why does the sum score prevail? Taking an examination is an intense experience, so that both instructors and examinees naturally favour a scoring system that is easy to understand and apply, is seen to be “fair” in some sense, and appears to do a good job of ranking students in a way that is consistent with other indications of academic success. Examinees seldom see plots like these two figures, and consequently tend to assume that failing items is their fault. What seems to be missing in the argument for more sophisticated analyses is a demonstration they would yield valuable benefits,

including the capacity to automatically correct the performance measure for problem items.

This paper takes advantage of a lack of identifiability in the modern test theory model to construct performance estimates $\hat{\theta}_j$ that resemble sum scores, but are actually optimally weighted sum scores that have substantially better bias and mean square error. We show that these $\hat{\theta}_j$'s automatically adapt to ICC shapes so as to resist the sources of sum score bias that we see in tests such as that in Figure 2. Our simulations suggest that these optimized scores would improve the performance of scoring for even designed tests such as the NAEP and SweSAT to an extent that would have socially important consequences for large scale testing.

First, we review the nature of the mathematical object $P_i(\theta)$ and clarify its connection to the data in Section 1.2; and also show that it is infinitely transformable. This discussion follows closely the more detailed treatment of Rossi, Wang and Ramsay (2002). Then we show that this lack of identifiability is, in fact, an asset that permits us to construct a performance measure that is designed to provide a side-by-side comparison with the sum score. In Section 2 we note that maximum likelihood estimation is explicitly a weighted sum measure with the weights depending on the ability level of the examinee. We develop a model for their distribution that generalizes the beta distribution. Then we apply a new technique called *parameter cascading* for fast computation of ICC's and their associated performance estimates. Finally, we show by simulations and fine-grained analyses of individual item response patterns how much improvement modern test theory makes possible.

1.2 The Performance Curve $\mathcal{P}$ and Model Identifiability

The value of an ICC $P_i(\theta)$, or of the ICC n -dimensional vector $\mathbf{P}(\theta)$, is probability and, as such, vector $\mathbf{P}(\theta)$ defines a space curve in the unit hypercube of dimension n where each axis is the probability of a correct answer on a specific item, and where the data for that item are positioned at the 0 and 1 endpoints. In the terminology of differential geometry, the test theory model defines a one-dimensional *manifold* that can be called the *test performance curve*, which we denote by $\mathcal{P}$.

But the $P_i(\theta)$'s are not functions in the usual sense, since the values of θ are not measured elements of a fixed set; but, instead, are transformable by any order-preserving function $h(\theta)$. That is, if $\xi = h(\theta)$ is a strictly monotonic transformation, then there exists a set of functions P_i^* such that

$$P_i(\theta) = P_i^*[h(\theta)] = P_i^*(\xi) \quad (1)$$

for all θ and all i . We shall also assume that h is differentiable.

This lack of identifiability does not depend on whether the ICC's are members of a low-dimensional parametric family or are represented within an arbitrarily flexible basis function system. As an example, consider the two-parameter logistic (2PL) model for $P_i(\theta) = \exp(a + b\theta) / [1 + \exp(a + b\theta)]$. It is pure convention that we express the model in this way rather than as $P_i^*(\nu) = a^*\nu^{b^*} / [1 + a^*\nu^{b^*}]$ where $\nu = e^\theta$. What are invariant under transformations such as the exponential are the 0 and 1 left and right asymptotes, respectively, the single inflection point, monotonicity, and differentiability. Of course the revised item difficulty parameter a^* and item discrimination parameter b^* have different algebraic roles than their (a, b) counterparts, but this is inconsequential as far as the capacity of the model to fit the data is concerned.

1.3 Historical remarks

Once, however, the convention is adopted that a particular low-dimensional model provides a satisfactory account of testing data, the shape of the θ distribution is resolved. Thissen and Orlando (2001) provide a nice account of the history of the now dominant 3PL convention. Thurstone (1925), probably because of the strong influence of the psychophysics of the time, proposed the standard normal distribution. Birnbaum (1968) noted that the difference between logistic functions and their normal ogive counterparts was too small to affect estimation, and that the logistic formulation is better for computation. But a formulation of IRT that produces scores distributed over the signed real numbers requires far more mathematical literacy than is available to most students, and to their teachers as well as a rule. Our goal is a formulation that is readily interpretable in terms of easily understood concepts such as the proportion of items correctly answered.

Nonparametric or functional ICC estimation has a substantial history in psychometrics. Mokken (1997) and others have implicitly considered nonparametric estimation in their studies of the implications of global ICC properties such as monotonicity on ability estimation. Ramsay (1991, 1997) used kernel smoothing over quantiles of the Gaussian distribution to provide fast and reasonably accurate ICC estimation in the program TESTGRAF. The θ metric was determined by assigning examinees to the N finite quantiles of the standard normal distribution, and then conditioning the estimation the ICC's on this assignment. TESTGRAF also implemented the expected sum score transformation of θ to $(0, n)$.

Rossi, Wang and Ramsay (2002) and Ramsay and Silverman (2002) optimized penalized marginal likelihood using the EM-algorithm, and as their roughness penalty increased, the estimates converged to something close in shape to the three-parameter

logistic model. Our method is different from theirs in several aspects, including replacing their use of the EM algorithm by a much faster approach, as well as replacing the entire real line for representing θ by $[0, T], 0 < T < \infty$, where two common choices for T are the number of items n and 100. Lee (2007) compared a number of nonparametric approaches. Woods (2006) and Woods and Thissen (2006) used a spline-based approximation to the ability distribution with simultaneous estimation of item parameters. Ramsay and Silverman (2005) developed methodology for nonparametric but strictly monotonic curve estimates; and, although we do not use that methodology in this paper, the importance of monotonicity for ICC estimation is clear and can easily be incorporated into our approach.

The next section defines how performance θ is estimated conditional on given ICC's. The estimation principle of parameter cascading is defined in Section 3. The fourth section contains a description of non-parametric ICC estimation. Section 5 describes the analyses of the two standardized tests and a classroom exam. Section 6 contains a simulation study and Section 7 provides discussion and conclusions.

2 The Estimation of Performance and the Item Impact Function

Any function $g(\mathbf{P})$ of the vector of the n correctness probabilities is invariant with respect to θ -transformations, and the expected total score $\sum_i P_i$ is one obvious choice of g . But expected sum scores inherit the illogicality of applying the same weight to all items, and we now explore a better idea and define a useful tool for representing score distributions.

2.1 Interpreting the conditional maximum likelihood equation

Let item function $W_i(\theta)$ be the log-odds ratio

$$W_i(\theta) = \log \left(\frac{P_i(\theta)}{1 - P_i(\theta)} \right). \quad (2)$$

The maximum likelihood estimate (MLE) $\hat{\theta}$ based on the 0-1 representations U_{ij} of wrong/right answers of examinee j to this item and conditional on previously estimates of each item's ICC, satisfies the equation

$$\sum_i^n [U_{ij} - P_i(\theta)] \frac{dW_i}{d\theta} = 0, \quad (3)$$

This relation is invariant with respect to strictly positive differential transformations $\xi = h(\theta)$; that is,

$$\sum_i^n [U_{ij} - P_i^*(\xi)] \frac{dW_i^*}{d\xi} = 0. \quad (4)$$

To see this, recall from (1) that the value of each P_i is invariant with respect to θ -transformations; and so, therefore, is any differentiable transformation of P_i , including the log-odds function W_i . By the chain rule, $dW^*/d\xi = (dW/d\theta)(d\theta/d\xi)$. Since by assumption $d\theta/d\xi > 0$, we may divide it into both sides of (3) to obtain (4). It must be pointed out, however, that it is the equation (3) that is invariant over θ -transformations, and not $dW/d\theta$ itself.

The importance of the W -derivative suggests that we give it a name, and we propose the *item impact* (II) function. How are we to interpret the optimal scoring equation (3)? Suppose, for a specific value of θ , we attempt to regress the residual n -vector $\mathbf{u}_j - \mathbf{P}(\theta)$ on the explanatory variable $d\mathbf{W}/d\theta$ of slopes of the W_i 's at value

θ . Then (3) is the numerator of the estimated regression coefficient. Consequently, at the MLE-optimal value of θ , these slopes will not have any explanatory power since this contribution is already present in the fit. The equivalent expression for the expected sum score is

$$\sum_i^n [U_{ij} - P_i(\theta_j)] = 0 \quad (5)$$

and we see that at the sum score value of θ log-odds slopes are apt to offer some improvement in the residual error sum of squares.

Both equations are, like the corresponding equation for the mean μ of a normally distributed sample, zero-valued sums of residuals and, as such, easy to explain to students in an introductory-level statistics class. They are inner products of the vector of residuals $U_{ij} - P_i(\theta)$ associated with ability θ , and the vector of either the item impact values for (3) or a vector of ones for (5). Each inner product is proportional to the correlation, defined relative to the base of zero for each factor, between the data/probability residuals and the item-impact values at θ , and the optimal value of θ is the point at which these two vectors are orthogonal; that is, where the correlation between the two factors is zero and therefore where the right item-impact vector offers no information in a linear sense about the residual vector. Since an item is informative about the value of θ when both $P(\theta)$ and $W(\theta)$ have a nonzero slope at that value, the item is treated as uninformative and the corresponding responses to the item for examinees at or near that performance level will have little or no impact on their estimated $\hat{\theta}$'s if an ICC and the corresponding W are flat. Moreover, if the slope is negative, as it is for the bright students for items 7, 58, 96 and 99 in Figure 2, then getting the item wrong will actually boost their $\hat{\theta}_j$'s. It therefore seems clear that (3) is to be preferred to the sum scoring equation (5).

The slope of the log-odds function, $dW(t)/d\theta$, also has a nice information theory

interpretation. For an event x with probability $P(x)$, $S(x) = -\log_2 P(x)$ is called the *surprisal* or *self-information* of the event. Surprisal takes only non-negative values, and has as its unit the bit since it is equal to the average number of successive coin-tossed heads whose probability is P (4.3 for $P = 0.05$, for example). Like magnitudes in the physical sciences, surprisals can be added and subtracted. The expected value of surprisal is *entropy* (Cover and Thomas, 2006).

If we use 2 as the base of the logarithm in computing $W(\theta)$, we see that

$$W(\theta) = S[1 - P(\theta)] - S[P(\theta)], \quad (6)$$

and therefore $W(\theta)$ has is the balance of information in favour of a successful outcome on an item for an examinee with ability θ : A large positive value implies a large surprisal for failure, zero implies probability 0.5, and a large negative value corresponds to a large surprisal for success. The slope of $W(\theta)$ is the instantaneous change in this balance at θ , measured in units *bit*/ θ . We use base 2 logarithms for all of our displays in this paper.

Figure 3 displays the estimated log-odds functions $W_i(\theta)$ for the NAEP and SweSAT in the upper panels along with their derivatives in the lower panels. These would be linear if the 2PL model represented all items, but the positive lower asymptote exhibited in the 3PL version implies a positive curvature in W_i . While there are plenty of curves in each test that have this shape, there are also many that do not. This is especially clear in the lower panels, where many items show item impact curves that change levels by large amounts as they move from the weakest students to the strongest.

Insert Figure 3 here.

The optimal scoring equation (4) implies two principles that are built into the

ability estimation methodology found in Section 3. First, the curve for an item that we really need is not $P(\theta)$, but rather $W(\theta)$, its log-odds transform. Second, whatever the approach to ICC estimation, there must be enough flexibility combined with enough smoothness in the process to track how the slope of $W(\theta)$ varies. Existing parametric models, including the 2PL/3PL family, cannot do this. These two conditions also collaborate to point to the use of spline functions as the solution to the estimation problem: first because splines, being linear combinations, are unbounded, secondly because they can assume arbitrarily complex shapes, and thirdly because constraining them to be smooth is a routine part of spline curve fitting technology. Add to this the fact that we are defining these curves over a closed interval and we have a context perfectly suited to splines.

2.2 The tilted scaled beta (TS β) distribution

Because θ can be transformed at will, we can define the range of θ to be the closed interval $[0, T]$, where T is a fixed upper limit such as 100 for percentage scoring. We envisage, at least for multiple choice items, that it is possible for an examinee to be either sufficiently limited or willfully ignorant, or to be sufficiently brilliant or lucky as to produce sum scores of 0 or T , respectively. We assume that sum scores of 0 or n must be assigned values of $\theta = 0$ or T , respectively.

We use a simple parametric model for the distribution of ability that we call the *tilted scaled beta* or TS β distribution to fix the distribution of θ and therefore identify uniquely the ICC's. Its probability density function is

$$p(S|\alpha, \beta, h_0, h_T) = \frac{h_0(1 - S/T) + h_1(S/T) + (S/T)^{\alpha-1}(1 - S/T)^{\beta-1}}{T(h_0 + h_T)/2 + B(\alpha, \beta)} \quad (7)$$

where $B(\alpha, \beta)$ is the beta function defined by parameters α and β , which in practice

will have a lower bound of two in order to ensure uni-modality. The $\text{TS}\beta$ density has the shape of the beta density, is scaled to be over $[0, T]$, but departs from beta shape in having the possibility of positive end-point heights h_0 and h_T . We optimize the $\text{TS}\beta$ log-likelihood with respect to the transformed parameters $\alpha^* = \log(\alpha - 2)$, $\beta^* = \log(\beta - 2)$, $h_0^* = \log(h_0)$ and $h_T^* = \log(h_T)$ and obtain rapid convergence using the Newton-Raphson algorithm.

Figure 4 shows the fits of the $\text{TS}\beta$ distribution to the sum scores of the NAEP and SweSAT. The NAEP requires its capacity to be positive at the score limits, and the distribution does a fine job of represent both distributions.

Insert Figure 4 here.

We suggest that schools and university departments might use the prior density $\text{TS}\beta$ to define targets for the distribution of test scores. For example, the four parameters of the distribution can be optimized to produce desirable proportions of the numbers 1 to 4 or the letters D to A often used grade point averages, so as to offset the grade inflation that has been apparent recent decades. The estimation methodology in Section 4 provides a method for smoothing the distribution of optimal scores toward such a reference density, and therefore provides an approach to norm-referenced testing.

The $\text{TS}\beta$ density is unimodal and therefore cannot be used for situations involving multimodality or more generally mixtures of populations having quite different ability distributions. There is a large literature on more general parametric and nonparametric representations of univariate density that can be consulted if this density is seen to be not appropriate (Ramsay and Silverman, 2005).

3 Parameter Cascading or Generalized Profiling

We employ an approach that has many similarities to marginalization over θ , but that leads to much faster computation and easier implementation for nonlinear models such as those used in analyses of testing data. Parameter cascading, abbreviated to PC in this paper, can be viewed as a generalization of *profiling*, often used in nonlinear least squares models (Bates and Watts, 1988; Murphy and van der Vaart, 2000; Severini and Wong, 1992). Profiling is used when a subset of the parameters, say θ , has a closed-form expression for the minimizer of a fitting criterion F given the remaining parameters in vector γ . That is, there is a functional relationship $\theta(\gamma)$ that can be cheaply computed. The optimization proceeds using the reduced dimension problem of numerically optimizing error sum of squares for γ by calculating the gradient using the total derivative

$$\frac{dF}{d\gamma} = \frac{\partial F}{\partial \gamma} + \frac{\partial F}{\partial \theta} \frac{d\theta}{d\gamma}. \quad (8)$$

PC imposes *regularity* or smoothness on the functional relationship $\theta(\gamma)$ so as to reduce the tendency of unrestricted estimates to over-fit the data, and to borrow strength from neighboring estimates by shrinking or smoothing them toward some reasonable target. A general strategy for achieving this in the nonlinear least squares context is to modify SSE to penalized least squares $SSE + \lambda PEN(\theta)$ where the penalty PEN term captures some form of departure from unrestricted variation and smoothing parameter λ controls the relative emphasis on this penalty. See Ramsay and Silverman (2005) and Ramsay, Hooker and Graves (2009) for extended discussions of regularization in the context of functional data analysis.

In PC the usual single-criterion optimization is replaced by a compound optimization involving an inner optimization of a criterion $H(\theta|\gamma)$ each time an outer criterion $F(\gamma)$ is improved by some perturbation of γ . The functional relationship

is implicit in this two-layer optimization. Either F or H may involve a negative log likelihood or any other appropriate loss function, and either may have additional roughness penalty terms.

By contrast, the *alternating* optimization (AO) often used in psychometric high dimensional models involves switching between optimizing F with respect to γ keeping θ fixed, and optimizing H with respect to θ keeping γ fixed. The linearly convergent EM algorithm is often used as an estimation algorithm for high dimensional parameters, as for example by Rossi, et al (2002). Like AO, it alternates between optimizations with respect to subsets of variables, but with the benefit that the E-step involves, in effect, a quadratic loss function having an explicit optimum. But, like PC, the two optimizations involve different criteria since the M-step is a maximum likelihood estimation.

The main difference between PC and AO/EM is in how the gradient and hessian are computed for the outer or top level function F . When the inner function does not have an explicit solution, the θ -derivative $d\theta/d\gamma$ in (8) is obtained via the Implicit Function Theorem as follows. Assume that the hessian $\partial^2 H/\partial\theta^2$ is positive definite at the optimizing value of θ conditional on γ . Then, because $\partial H/\partial\theta = 0$,

$$\frac{d}{d\gamma} \frac{\partial H}{\partial\theta} = \frac{\partial^2 H}{\partial\theta\partial\gamma} + \frac{\partial^2 H}{\partial\theta^2} \frac{d\theta}{d\gamma} = 0 \quad (9)$$

from which we obtain

$$\frac{d\theta}{d\gamma} = - \left[\frac{\partial^2 H}{\partial\theta^2} \right]^{-1} \frac{\partial^2 H}{\partial\theta\partial\gamma} \quad (10)$$

and the gradient or total first derivative of F with respect to γ is

$$\frac{dF}{d\gamma} = \frac{\partial F}{\partial\gamma} - \frac{\partial F}{\partial\theta} \left[\frac{\partial^2 H}{\partial\theta^2} \right]^{-1} \frac{\partial^2 H}{\partial\theta\partial\gamma}. \quad (11)$$

Alternating optimization, on the other hand, only uses the first term $\partial F/\partial \gamma$ in this expression. For this reason, alternating optimization methods exhibit only linear convergence, whereas the PC algorithm converges quadratically using the quasi-Newton or Newton-Raphson methods, and hence typically takes many fewer iterations.

4 Parameter Cascaded ICC Estimation

We now apply the PC approach to the modeling of binary testing data. We include a roughness penalty in both the inner and outer optimization criteria. These regularization terms exchange some bias in a parameter estimate in return for a reduction in sampling variance large enough to result in an interesting reduction in mean square error (MSE), where $\text{MSE} = \text{Bias}^2 + \text{Var}(\hat{\theta})$. For example, the statistic $U_j = \sum_i U_{ij}$ is an unbiased estimate of $E(U_j) = \sum_i P_i(\theta(j))$ but we hope, by applying a more sophisticated estimation scheme, to reduce the sampling variance of this elementary statistic enough to achieve an interesting improvement in the MSE of a θ estimate.

4.1 The Expression of $W_i(\theta)$ as Spline Functions.

The approximation of functions in terms of linear combinations of basis functions bring two advantages: basis function expansions are easy to expand in terms of dimensionality; and, given a good choice of basis system, they usually lead to faster and more stable computation. However, linear combinations do not respect constraints such as $0 < P_i < 1$, so that it is preferable to use linear expansions of the log-odds transformation,

$$W_i(\theta) = \sum_k^K \gamma_{ik} \phi_{ik}(\theta) = \gamma'_i \phi_i(\theta), \quad (12)$$

where γ_{ik} is the coefficient of basis function ϕ_{ik} in the basis function expansion of the i th ICC. We use the B-spline basis functions $\phi_k(\theta) = B_k(\theta|\boldsymbol{\xi}, M)$, where $\boldsymbol{\xi}$ is a knot sequence and M is the order of the spline.

A challenge for all test theory models is the sparsity of information about the ICC's for extreme values of θ , so that estimates of curve values over these regions are inevitably much less stable. This argues against the usual practices in spline smoothing of either using a uniform knot sequence, or positioning a knot at each value of the independent variable. Instead, we obtained much better results by positioning knots so as to have roughly equal proportions of sum scores for each inter-knot interval. This is easily achieved by positioning knots at a set of $k - M$ equally-spaced interior quantiles of the $\text{TS}\beta$ distribution fit to a particular score distribution, where k is the number of spline basis functions and M is the order of the spline.

4.2 Initialization of the optimization.

This knot selection strategy also leads to a fast, stable and remarkably efficient initial estimates of the coefficients γ_{ik} . Each of the $k - M + 1$ within-interval average binary scores is converted by the log-odds transformation to corresponding values of the W -functions, and these are then smoothed using a standard spline smoothing algorithm as is found in the functional data analysis R package `fda`. Because the data in this smoothing operation are log-odds transformations of proportions, they have sampling distributions that do not vary a great deal from one set of data to another with a fixed sample size. Thus the smoothing parameter λ is primarily a function of N . For the two data sets in Figure 1, where $N \approx 2500$, $\lambda = 100$ defined estimated W -functions that were smooth but with enough detail in their curvatures to defined plausible variations in the shapes of the corresponding ICC's. Finally, initial values for the

θ_j 's can be determined quickly by identifying the integer values minimizing the inner criterion H , given these initial ICC estimates, although the scores themselves also serve well as initial estimates of the θ_i 's.

4.3 The inner optimization criterion $H(\boldsymbol{\theta})$

The inner optimization criterion defines the loss function H for fitting θ to an examinee's binary performance vector $\mathbf{u}_j = (U_{j1}, \dots, U_{jn})$ given the current state of the structural parameters $\boldsymbol{\gamma}$. We use for this purpose the negative log likelihood

$$-\log L(\boldsymbol{\theta}|\boldsymbol{\gamma}, \mathbf{U}) = \sum_{i=1}^n \sum_{j=1}^N [U_{ji} W_i(\theta_j|\boldsymbol{\gamma}_i) - \log[1 + \exp W_i(\theta_j|\boldsymbol{\gamma}_i)]] \quad (13)$$

where $\mathbf{U}$ is the N by n binary data matrix.

However, H must also define a relatively smooth relationship $\boldsymbol{\theta}(\boldsymbol{\gamma})$, where "smooth" is defined in a way that is appropriate to the problem. This is achieved by appending a penalty term to the negative log data-likelihood as follows:

$$H(\boldsymbol{\theta}|\boldsymbol{\gamma}, \mathbf{U}) = -\log L(\boldsymbol{\theta}|\boldsymbol{\gamma}, \mathbf{U}) - \lambda_H \sum_j^N (\theta_j - Q_j)^2 \quad (14)$$

where Q_j is the corresponding quantile of a $\text{TS}\beta$ distribution. As smoothing parameter λ_H increases, the estimated θ_j 's are forced to approach these quantiles, so that when $\lambda_H \rightarrow \infty$ the data in effect estimate the quantile of the $\text{TS}\beta$ distribution corresponding to θ .

We optimize H using Newton-Raphson iterations, which can be vectorized by noting that the solution for each θ_j depends only on $\mathbf{u}_j$, the vector of N scores for item i ; and these N single variable optimizations can also profit from the easy availability of expressions for both the gradient and hessian. Typically, only small

changes are made in γ in the outer optimization, resulting in the need for only a few inner optimization iterations each time the coefficients γ_{ik} are updated, so that the computational burden at the inner level is only a few times greater than if an analytic expression were available for the optimal values.

4.4 The outer optimization criterion $F(\gamma)$

The outer criterion F also uses the same data-fitting negative log likelihood, but this time it is viewed as a function of the coefficients in γ so as to optimize the shapes of the functions W_i . Since it is a function of γ both directly and through the relation $\theta(\gamma)$, the gradient calculation (11) requires the second cross-derivative of criterion H with respect to both γ_i and θ_j . In order to ensure that the estimated ICC's are smooth when larger numbers of basis functions are used, the roughness penalty using second derivatives that is often used for piecewise cubic B-splines can be appended:

$$F(\gamma|\mathbf{U}) = -\log L(\gamma|\mathbf{U}) + \lambda_F \sum_i^n \int_0^T \left[\frac{\partial^2 W_i}{\partial \theta^2} \right]^2 d\theta. \quad (15)$$

The penalization of the integrated squared second derivative implies that as $\lambda_F \rightarrow \infty$ the W_i 's will approach straight lines, which characterize the two-parameter logistic model. The use of the third derivative would instead use quadratic polynomials as the target for smoothing. More sophisticated options for linear differential operators can be considered, as illustrated in Rossi, et al. (2002) and discussed in Ramsay and Silverman (2005).

4.5 A summary of the algorithm

The algorithm that we have used and that we propose has these six steps:

1. Select or estimate a $\text{TS}\beta$ model.
2. Compute $K - 1$ interior quantiles of the $\text{TS}\beta$ distribution to define the internal boundaries between K score bins. These bins will contain roughly equal numbers of the sum scores. Compute the log-odds transformation of the proportion of correct answers in each bin.
3. Compute initial estimates of the ICC's by smoothing the binned data as a function of score values at the bin-middles using $K + 3$ order four B-spline basis functions with knots located at the bin boundaries. After some experimentation, we have found a smoothing parameter of $\lambda = 1000$ to be about right, and this value should serve reasonably well for most tests.
4. Compute the initial θ estimates conditional on the smoothing estimates of the ICC's, which we refer to as CS-MLE estimates. We selected the best MLE value over a fine grid; but, if preferred, one can use a one-dimensional optimization strategy.
5. Using the ICC coefficients and the CS-MLE θ values as initial values, optimize ICC estimates using PC. Our simulation results, described in Section 6, indicate that going all the way to convergence gains little over stopping at about 20 iterations.
6. Generate the final un-penalized PC-MLE theta estimates using smoothing parameter $\lambda_H = 0$.

This algorithm will analyze tests with samples of the sizes that we used in this paper in a few seconds on most computers. We used a quasi-Newton optimization algorithm for all of our analyses.

5 The analysis of three sets of test data

Since the NAEP and SweSAT had much larger number of examinees than the psychology test, we describe the two analyses separately. Recall that the initial θ scores estimated by maximum likelihood using the initial ICC's estimated by binning the data are the CS-MLE scores, and the final converged ICC-based scores as PC-MLE scores.

5.1 The NAEP and SweSATs

Since these designed tests, we focus on the test-developer's perspective that it is the ICC's that are the main focus of the analysis. Figure 4 shows the distribution of the sum scores for the two tests. We see that the NAEP data contain a substantial number of 0 and n scores; but that SweSAT data have scores ranging from only 20 to 118.

The first step in the analysis of each data was the maximum likelihood estimation of the $TS\beta$ distribution for each set of sum scores. We do not advocate this choice of reference distribution in general, but in Section 6 we compare PC results to sum score results using populations defined the results of these analyses. We wanted to be as accommodating as possible to sum scores in order to provide a fair comparison, and therefore chose here to smooth the score distribution to that of the sum scores. The values of α were 4.30 and 5.61 for the NAEP and SweSAT data the corresponding β values were 3.45 and 4.61, respectively; indicating a more peaked distribution for the SweSAT. The h_0 values were 0.00066 and 0.00000, and the h_T values were 0.00138 and 0.00001, respectively. Figure 4 displays the respective density functions re-scaled so as to overlay the sum score frequencies, and we see that the NAEP data imply rather larger values of the endpoint heights. The $TS\beta$ densities do a fine job of

approximating the respective sum score distributions.

We chose $k = 10$ order 4 B-spline basis functions for representing the log-odds functions W_i for both sets of data in order provide enough flexibility to represent ICC features at a level of detail that the large samples sizes make possible. The six interior knots were positioned so that roughly equal proportions of sum scores occupied the inter-knot intervals. Initial values for both levels of optimization were generated by applying the smoothing algorithm to estimate values of the W -functions, followed by calculating the maximum likelihood estimates of θ conditional on these ICC curves. The smoothing parameter values for the optimization of the ICC estimates were $\lambda_H = 0.01$ and $\lambda_F = 0.1$, determined by the using cross-validation process and also by inspecting the results. The NAEP and SweSAT optimizations required 70 and 7 iterations on our computer, and took 27 and 6 seconds, respectively.

The ICC's are displayed in Figure 1. Most of the NAEP curves conform well to the classic ogive represented by the 3PL model, although a few have a plateau or staircase pattern. The shapes of the SweSAT curves are more variable, often departing substantially from the ogive, and frequently not approaching one on the right. The corresponding W -functions are shown in Figure 3.

Insert Figure 5 here.

Although the prior density for the performance estimate $\hat{\theta}$ is the $TS\beta$ density function (7), the actual estimates by maximum likelihood can systematically deviate from the sum scores as the data and the model require. Figure 5 plots the differences between the PC-MLE estimates of θ_j and the sum scores against the corresponding sum scores for both tests. For the NAEP we see that the lower 25% gains about one item on the average, the median examinees lose on average about one item, and the top 5% of the students gain as much as four items. The overall effect is a more peaked

density over the central-high region accompanied by rather longer tails, especially at the top end. The same pattern is seen for the SweSAT, except the score gains and losses are larger because the test is longer.

The test information functions for the two tests are shown in Figure 6, and we see that the NAEP is most informative at the 25% level while the more challenging SweSAT peaks at the 75% score.

Insert Figure 6 here.

Since the PC approach defines abilities θ_j as functions of the ICC parameters γ_{ik} , we can use cross-validation to estimate the roughness penalty parameters λ_H and λ_F . For each test, we randomly assigned 1600 examinees to the training sample, and the remainder to the validation sample. For the NAEP the best validation sample penalized log likelihood was achieved for $\lambda_H = 0.01 = \lambda_F = 0.01$, and for the SweSAT for $\lambda_H = 0.001$ and $\lambda_F = 0.01$.

5.2 The psychology test

We now move to a focus on the test-taker, since this was a one-time classroom test administration where, as far as we know, the test designer did not do any item screening. Figure 2 shows us that the sum score is not a viable benchmark against which to assess the performance of a maximum likelihood estimate of either a $P_i(\theta)$ function or a performance level θ_j since the number of faulty or uninformative items is sure to lead to considerable bias. Rather, the focus here is on MLE as a robust estimation technique that can offset the impact of these items and present a better picture of each examinee's actual command of the course.

We analyzed these data using the same optimization settings that we employed for the two industrial tests except that, because of the smaller sample size, we used eight

instead of ten knots. The values of α and β were 9.1 and 5.1, respectively, the end-point heights were both zero. Figure 7 plots the differences between the optimal and sum scores in scores against the sum scores. For the optimal scores, we see that the examinees in the 25% to 95% range lose as many as 10 items due to the discounting of the scores on the extremely easy items. But the top students gain a great deal since their sum score penalty for answering the descending probability items “incorrectly” switches to a positive increment. The top sum score was 91 and this was replaced by 99, which seems like a more likely score for a highly competent student on a multiple choice test in an introductory level course. The median score decreases by about three items.

Insert Figure 7 here.

6 Simulated data experiments

In this section we examine the performance of the PC algorithm for simulated data resembling what we would see from multiple administrations of the NAEP, SweSAT and psychology tests, each to a new set of examinees. Each simulated sample was generated using as population ICC’s the curves for one of the three real data samples, and as population abilities random θ ’s drawn from the associated estimated TS β distribution fit to each test’s PC-MLE $\hat{\theta}$ distribution.

The main goal was to compare the quality of the estimates of the ICC’s and θ ’s computed from sum scores to that of the PC estimates. In order to give sum scores as much advantage as possible, we used as the population distributions of ability the respective TS β density functions estimated from the sum scores in the previous Section, even though this tends to bias our results in favour of sum scores.

We also wanted see how the two types of regularization, corresponding to the two smoothing parameters λ_H and λ_F , affect the quality of the estimated W -functions. Each sample was analyzed using each of the 49 smoothing parameter pairs (λ_H, λ_F) produced by the pairing the values $\log_{10} \lambda = -3, -2, -1, 0, 1, 2$. We used samples size N of 200, 400, 800 and 1600. That is, the experimental design is between-treatment with respect to N and test, and within-treatment with respect to (λ_H, λ_F) . Five hundred simulation samples were analyzed for each sample size and for each test.

The analysis of each simulated sample was carried out in exactly the same way as the analysis of the corresponding real data sample, except for the (λ_H, λ_F) pairs. The same number and order of B-spline basis functions and knot choice procedure was employed, and knots were assigned uniquely for each simulated sample.

6.1 Quality of estimation of $W(\theta)$

The log-odds functions $W(\theta)$ are more directly involved in the test analysis than the ICC's themselves and, in addition, have unconstrained values. These aspects argue for assessing ICC estimation quality in terms of these functions.

The two measures of ICC quality were measures of the bias and root mean squared error associated with the estimated W -functions. These were

$$\begin{aligned} \text{Bias} &= \frac{1}{500n} \sum_{\ell}^{500} \sum_i^n \int_0^T [\hat{W}_{i\ell}(\theta) - W_i^*(\theta)] p_{\ell}(\theta) d\theta \\ \text{RMSE} &= \sqrt{\frac{1}{500n} \sum_{\ell}^{500} \sum_i^n \int_0^T [\hat{W}_{i\ell}(\theta) - W_i^*(\theta)]^2 p_{\ell}(\theta) d\theta} \end{aligned} \quad (16)$$

where $W_{i\ell}^*$ and $\hat{W}_{i\ell}$ are the population and estimated functions, respectively, for item i and simulated sample $\ell = 1, \dots, 500$. The probability density p_{ℓ} in the integration was that of the TS β distribution estimated for each sample, and this weighting was

used to reduce the influence of the unstable estimates of W for extreme abilities that would result from using uniform measure, and tend to dominate the RMSE value.

The results for the estimation of the ICC's are in Table 1. The root mean squared errors for the optimized ICC's are slightly higher for the NAEP than the corresponding values for the SweSAT for all values of N , reflecting that fact that, in total, the SweSAT contains more information. For illustration purposes, an $\text{RMSE} = 0.2$ associated with the log-odds value $W = 0$ translates to a shift in the probability value $P = 0.5$ of 0.05, or a 10% error. This seems quite satisfactory, and even corresponding RMSE's for $N = 200$ suggest quite useful estimates of ICC's for samples that small.

Insert Table 1 here

We were impressed with the quality of the initial binned smoothing values in the second column, which involve only a roughness penalty on spline-smoothing estimate of W . The final column suggests that the optimized PC only achieves about a 6% improvement over these initial estimates for $N = 800$ and 1600. It appears that the rapidly computed initial smoothing estimate of the ICC's would be quite adequate when the test analysis is more exploratory in nature. However, the improvement achieved by optimizing is rather larger for the SweSAT when smaller N 's are involved.

It is also clear, as expected, that smoothing the θ -estimates toward the more stable $\text{TS}\beta$ quantiles paid off in terms of better ICC estimates. The first column of RMSE values shows that using no regularization in the inner optimization produces ICC estimates that are worse than even the initial estimates for all except the largest sample sizes. The amounts of the two regularizations does vary between the two tests, suggesting that shorter tests like the NAEP will need a higher penalty for θ and a lower penalty for W . This is a consequence of shorter tests yielding noisier estimates of θ .

Not shown in Table 1 are the corresponding bias results. These were so small in all conditions that their contribution to RMSE could be considered negligible. Also not shown are results for small fixed numbers of iterations in the optimization. We consistently observed that twenty iterations yielded ICC qualities indistinguishable from those in the table for the converged estimates,

6.2 Quality of estimation of θ

Table 2 displays the RMSE values for three estimates of the values of θ_j used to generate the simulated data for the two designed and one classroom test. The estimates are (1) the sum score, (2) the CS-MLE estimates used to initialize the PC optimization, and (3) the final PC-MLE estimates. Also shown is the percent reduction in the RMSE for the sum score achieved by the lower of the other two estimates.

Insert Table 2 here

Overall, the RMSE values do not vary much between the $N = 400$ and $N = 1600$ sample sizes, so that the noisier ICC estimates for $N = 400$ seems not to handicap θ estimation significantly. The sum scores have substantially higher RMSE's, primarily due to their large bias as estimates of $E[\sum_i U_{ij}]$ rather than θ_j . But, even for the two tests with items selected to favour the sum score as an ability estimate, an improvement in ability estimation over sum scoring is achieved that would have valuable consequences if installed on a large scale given that nearly every citizen is tested at some point in life.

Not surprisingly, the roughly 14% improvement over sum scores for the classroom test was much larger. What is especially valuable from an application perspective is how efficient the cheaply computed CS-MLE estimates are. A comparison with

Table 1 suggests that good estimates of θ does not need a high level of accuracy in the estimated ICC's.

Figure 8 provides a more detailed picture of how bias and RMSE vary across vary over θ . The greatest beneficiaries of the complete PC-MLE algorithm are the top students. The bias in the NAEP at highest score is about -1 item for PC-MLE, but more than three times as large for both the sum score and CS-MLE estimate. Analogous results for the SweSAT are -2 items and four times as large otherwise, and for the psychology test -2.5 items and also four times as large for sum score and CS-MLE. Similarly, RMSE's at the high end for the NAEP are 2, 3 and 3.8 items for PC-MLE, CS-MLE and sum scores, respectively, for the SweSAT about 3, 9 and 9 items, respectively; and for the psychology test about 4, 10 and 10 items, respectively. The fact that these top examinees also benefit relative to the CS-MLE scores justifies the extra computational effort for tests as important as the NAEP.

All three tests exhibit relatively small bias for the central 50% of the scores, but the sum score has a RMSE of about 2.6 items whereas the two MLE scores are at about 2.25 items or less because it does not take advantage of the information provided by the slopes of the log-odds functions.

Insert Figure 8 here.

7 Discussion and conclusions

Our principal conclusion follows from our analysis of the conditional maximum likelihood stationary equation (3) and is illustrated in the lower panels of Figure 3. This is that a differential weighting of items where the weights depend on ability level as well as items can markedly improve the estimation of ability, and especially so for

some score groups such as very high performing examinees. That is, the weighting systems depends on an *interaction* between item and ability.

Once this conclusion is accepted, we can envisage many scoring schemes that would take advantage of this insight, and we have presented two: a simple and fast preliminary smoothing estimation of the log-odds functions W_i , and a more up-market optimal estimation of these functions with an implicit estimation of abilities based on the principal that ability estimates are best viewed as nuisance parameters and the W_i as functional structural parameters. We are sure that other approaches will emerge.

A second conclusion is that recasting item response theory into the more familiar framework of ability estimates varying over closed non-negative intervals such as $[0, n]$ or $[0, 100]$ will aid the acceptance of weighted scoring schemes within evaluation contexts where the sum score now prevails. To this end we have taken care to explain why the performance manifold $\mathcal{P}$ is effectively the model for the data; and as such, is identified by the data rather than being latent. Ability is a metric laid down along the performance manifold space curve that has the continuous, connected order topology of the real line, but is otherwise arbitrary. Many if not most texts on test theory argue that the theory can define an ability metric that can be used to compare examinee performances on the basis of different tests and different examinee populations. Equation (1) should appear early in future texts in order to dispel that impression.

Some choice among this infinitude of arbitrary metric possibilities must be made in order to work with the test theory model, and testing methodology has traditionally resolved this problem by agreeing to use a fixed algebraic family of item characteristic curves, most notably the two- and three-parameter logistic models. But we depart from this framework for two reasons. First, it is clear from the examples presented

in this paper that these models do not adequately represent the data; and second, that the more natural representation of the item characteristic is in terms of the more nearly linear and otherwise unrestricted log-odds function $W_i = \log[P_i/(1 - P_i)]$. Five-dimensional order four spline functions with a single interior knot close to the median score have worked well for us.

We propose as a useful family of models for ability distribution a four-parameter probability density function that can both capture the possibly nonzero density at score boundaries and the usual shape characteristics in the interior. The $TS\beta$ density works well in our experience and is easy to identify from data, but our methodology can be adapted to other parametric or nonparametric characterizations of score distributions as required. Although it might be tempting to anchor the ability density by using this model to fit the sum score distribution, we imagine that administrators of testing programs would consider pre-defining this density by either using the $TS\beta$ model for test equating or by using it to represent predefined target proportions for score categories such as the 0 to 4 and A to D systems often used for grade-point averages.

We argue that it is both intrinsic to the test theory model and evident in the nature of abilities θ_j versus item parameters γ_{ik} that the former should be viewed as smooth functions $\theta_j(\boldsymbol{\gamma})$ of the latter. The parameter cascading strategy that we use for computing estimates of both parameter classes is based on this principle. In effect, PC reduces the dimension of the parameter space from $nK + N$ to nK item parameters, about a 8 to 1 ratio in the NAEP case, and this makes possible the use of cross-validation for model identification and smoothing parameter selection. Substantial speed-up of computation is a useful byproduct, and confidence interval estimation and inference can use standard methods since the dimension of the parameter space does not depend on the number of examinees.

Our goal is to present and use test theory in a way that can be more easily understood by teachers and administrators in educational institutions and, even more importantly, by the examinees whose futures often depend critically on the characterization of their performance on tests. Our use of the closed $[0, n]$ interval combined with a metric resembling that of sum squares allows us to display, we believe for the first time, a quantitative assessment of advantages of using modern test theory over classical test theory for both designed and one-off classroom tests. The improvements in ability assessment in Table 2 are sure to seem important to both educational administrators and to the students for whom they are responsible, and especially so for the top performers.

The additional effort required can be made negligible through the provision of free or low-cost software to instructors, and we are currently working on applications to serve this end. We would also argue that there is much to gain for students by using the same software to display test ICC's; along with summary measures of the distribution of assessments, for which the $TS\beta$ is a nice tool.

Finally we acknowledge that achieving a wide penetration of more efficient processing of test data into the current testing culture is sure to be a slow and often difficult process requiring decades, as health-related interventions such as weight control and smoking cessation have already demonstrated. Students in particular will want to know which items had the most impact on their score, and are apt to want to see exam questions after test administration. Current classroom and designed testing practice does not encourage this. However, the great progress achieved in statistical graphics combined with a wider awareness of the statistical challenges involved are likely to be valuable assets.

References

- Birnbaum, A. (1968.) Some latent trait models and their use in inferring an examinee's ability. In F. Lord and M. R. Novick (Eds.) *Statistical Theories of Mental Tests*, pp. 395-479. Reading, MA: Addison-Wesley.
- Gulliksen, H. (1950) *The Theory of Mental Tests*. New York: John Wiley & Sons, Inc.
- Lee, Y-S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement*, 37, 121-134.
- Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden and R.K. Hambleton (Eds.), *Handbook of modern item response theory* (351-367). New York: Springer.
- Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95, 449-485.
- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation, *Psychometrika*, 56, 4, 611-630.
- Ramsay, J. O. (1996) A geometrical approach to item response theory, *Behaviormetrika*, 23, 3-16.
- 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., Cao, J. and Campbell, D. (2007). Parameter estimation for differential equations: A generalized smoothing approach (with discussion), *Journal of the Royal Statistical Society Series B*, 69, 741-796.
- 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 Silverman, B. W. (2005). *Functional data analysis. 2nd edition*, New York: Springer-Verlag.
- 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.
- Severini, T. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics*, 20, 1768-1802.
- Thissen, D. and Orlando, M. (2001) Item response theory for items scored in two categories. In D. Thissen and H. Wainer (Eds.) *Test Scoring*, pp 73-140. Mahwah, NJ: Lawrence Erlbaum Associates.
- Cover T. M. and Thomas, T. M. (2006). *Elements of Information Theory, Second Edition* Hoboken, NJ: Wiley.
- Thurstone, L. L. (1925) A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433-449.
- 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.

Table 1: The quality of estimation of log-odds functions $W_i(\theta)$ in terms of root mean squared error measure (16). The percent improvement in the last column is in the optimal RMSE relative to the initial RMSE.

N	$\lambda_H = \lambda_F = 0$	Initial	Optimal	Optimal λ 's	%
NAEP	$RMSE \times 1000$			$(\log_{10} \lambda_H, \log_{10} \lambda_F)$	Improvement
200	430	406	383	(0, 0)	5.7
400	319	298	282	(1,-2)	6.0
800	240	226	213	(1,-2)	5.8
1600	180	178	166	(1,-2)	6.7
SweSAT					
200	490	415	353	(-3,1)	14.9
400	367	290	258	(0,0)	11.0
800	290	212	194	(-1,0)	8.5
1600	160	160	151	(-1,0)	5.6

Table 2: Root mean square (number of items) for the estimation of ability θ_j for three tests. CS-MLE is the initial estimate of θ produced by smoothing binned data and computing the conditional maximum likelihood estimate. PC-MLE is the estimate resulting from PC. The percent change is from the sum score to the better of the two alternatives.

N	Method	NAEP	SweSAT	Psych
	Sum Score	2.61	5.09	5.08
400	CS-MLE	2.45	4.95	4.59
	PC-MLE	2.50	4.96	4.42
	% change	5.1	2.1	12.8
1600	CS-MLE	2.48	4.81	4.53
	PC-MLE	2.45	4.79	4.35
	% change	6.5	5.7	14.4

Figure 1: The left panel displays estimated ICC's for a 36-item National Assessment of Educational Progress (NAEP) history test, and the right panel displays these curves for a 122-item Swedish Scholastic Aptitude Test (SweSAT) interval. In each panel the vertical dashed lines represent the 5%, 25%, 50%, 75% and 95% quantiles, respectively, of the tilted scaled beta distribution (7) fit to the sum score distributions.

Figure 2: The left panel contains the item characteristic curves for a test of length 100 given to 379 students in introductory psychology at McGill University. The heavy lines identify six items that discriminate against the best students. Nine items are answered correctly by almost everyone, and are therefore also uninformative. The right panel displays the base 2 log-odds transformation $W = \log_2[(P/(1 - P))]$ of the curves in the left panel.

Figure 3: The upper left panel displays estimated logit-transformed ICC'Ss $W_i(\theta)$ for the NAEP, and the upper right panel displays these curves for the SweSAT. The lower panels display the derivatives of the upper curves, the item impact curves, that provide the optimal weighting of item scores.

Figure 4: The left panel displays the distribution of sum scores for the 36-item NAEP, and the right panel displays these scores for the 122-item SweSAT. In each panel the smooth curves are the densities of the tilted scaled beta distribution rescaled to represent frequencies.

Figure 5: The left panel displays the differences between the PC maximum likelihood estimates of the performances θ_j and the sum scores S_j plotted against the sum scores for the NAEP. The corresponding plot for the SweSAT is in the right panel.

Figure 6: The left panel displays the test information function for the NAEP history test, and the right panel displays the test information for the 122-item SweSAT.

Figure 7: The differences between the PC maximum likelihood estimates of the performances θ_j and the corresponding sum scores S_j are plotted against the sum scores for the psychology test. The top sum score of 91 increases to 99.

Figure 8: The mean bias (top) and RMSE (bottom) for the PC-MLE (solid), CS-MLE (dashed) and sum score (dotted) for the NAEP exam.