Relationship Between Students' Test Results and Their Performance in Higher Education Using Different Test Scores

Marie Wiberg, Juan Li, Per-Erik Lyrén, and James O. Ramsay

Abstract The aim of this study is to examine the relationship between students' college admissions test results and their performance in higher education using sum scores and optimal full-data scores. We used students from four university programs to examine predictive validity in terms of how the students performed on their studies in terms of obtained credits, as compared with their college admissions test results. The students' test results were calculated using the commonly used sum scores and the recently proposed optimal scores. We also examined the predictive validity of the test scores while controlling for the student backgrounds in terms of educational background, migration background, and gender. The results show that using optimal scores or sum scores yields slightly different test score distributions, especially the score distribution among the highest test performers differed. Practical implications of which test scores to use in college admissions testing in the future are discussed.

Keywords Optimal scores · Sum scores · Predictive validity · College admissions test

M. Wiberg (✉)
Department of Statistics, USBE, Umeå University, Umeå, Sweden
e-mail: marie.wiberg@umu.se

J. Li
Ottawa Hospital Research Institute, Ottawa, ON, Canada

P.-E. Lyrén
Department of Applied Educational Science, Umeå University, Umeå, Sweden

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

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 393, https://doi.org/10.1007/978-3-031-04572-1_4

1 Introduction

Sum scores are used in most standardized tests across the world, for example the Graduate Record Exam (GRE) (GRE, 2021), the Scholastic Aptitude Test (SAT) produced by the College Board, the aptitude tests produced by American College Testing (ACT) (Dorans, 1999), and the Swedish scholastic aptitude test (SweSAT) (Lyrén et al., 2014). Recently, optimal scoring was proposed as an alternative to sum scores which takes care of the different information provided by different items (Ramsay & Wiberg, 2017a,b). An advantage with optimal scores is that the scores becomes fairer in comparison to the actual knowledge level and thus high achievers may achieve higher optimal scores than sum scores as seen in Wiberg et al. (2019).

In Sweden, the selection component in admissions to university (the other component being eligibility) is based on a quota system. A certain proportion of candidates are admitted from quota groups based on different selection instruments, where grades from upper-secondary school (USGPA) and scores from the optional admissions test, SweSAT, are the two most common ones. Candidates who have both a valid USGPA (which most candidates have) and a valid SweSAT score will be placed in both quota groups, so taking the SweSAT can only increase one's chances of being admitted. If a test taker has several valid SweSAT scores, the best score is used in the admissions.

The goal of the SweSAT is to select those students who are most likely to perform well in higher education. Consequently, as is the case with any other selection instrument, the predictive validity of the scores is central to the overall validity of the use and interpretation of SweSAT test scores. Predictive validity studies on selection instruments in Sweden have often compared the predictive strength of the USGPA and SweSAT scores. The most common finding is that the USGPA is a better predictor than SweSAT scores (Svensson et al., 2001; Cliffordson, 2008; Cliffordson & Askling, 2006; Lyrén, 2008) and that the predictive strength differs between university programs for both instruments. For example, Lyrén et al. (2014) analyzed eleven different programs and found that the correlation between SweSAT scores and the performance criterion was non-significant for two programs (medicine and social work) and that it varied between 0.2 and 0.4 for the other nine programs (engineering, nursing, economics, teaching, etc.). They also found that the correlations were similar for the two section scores (Verbal and Quantitative), except for engineering programs where the correlations were higher for the Quantitative score than for the Verbal score.

In this paper we are interested in examining the predictive validity of the SweSAT if we use full information optimal scoring (Ramsay et al., 2020) as compared with using sum scores. Previous studies with optimal scores have focused on examining the possibility to use optimal scores instead of sum scores when we have binary scored multiple choice items (Ramsay & Wiberg, 2017b,a) and also a comparison between full information optimal score and binary sum scores (Wiberg et al., 2018) as well as a comparison between binary optimal scores and item response theory scores (Wiberg et al., 2019). In this paper we use full information optimal scores as

described in Ramsay et al. (2020). It is called full information optimal score because information in both correct and incorrect responses was used for scoring. This paper is different from previous papers as the focus is not to refine optimal scoring but instead to examine the predictive validity of the optimal scores in terms of how students perform once they have been admitted to a university program of their choice. The overall aim is to examine the predictive validity and thus the relationship between students' college admissions test results and the students' performance in higher education using sum scores and optimal scores.

The rest of this paper is structured as follows: Next, the method section with the different test scores, the sample used and the statistical analysis are described. This section is followed by a result section and the paper ends with a discussion with some concluding remarks.

2 Method

2.1 Test Scores

We focus on two different kinds of test scores; sum scores and full information optimal scores. For multiple choice items, sum scores are typically defined as the number of items the test taker answered correctly. The full information optimal score, further referred to as optimal scores, is formally defined in Ramsay et al. (2020) and thus only briefly described here. In the later empirical study, we use the freely available software R and especially the package TestGardener (Ramsay & Li, 2021) to estimate the optimal scores. The initial proposal of optimal scores were made by Ramsay and Wiberg (2017b).

The basic idea is to estimate the scores based on the interaction between the performance of items/options and test taker; and using surprisal $W_{im}(\theta) = -log_M P_{im}(\theta)$ rather than probability $P_{im}(\theta)$ in the estimation process, where θ is the given test taker's ability and M is the number of options of item i. Let $m = 1, \ldots M_i$ represent the different answer options for item i, and let $P_{im}(\theta)$ be the probability of a test taker with ability θ choosing the m option. The multinomial item response function can then be defined as

$$P_{im}(\theta) = \frac{\exp\left[W_{im}(\theta)\right]}{\sum_{l=1}^{M_i} \exp\left[W_{il}(\theta)\right]}. \tag{1}$$

where W_{im} is an unbounded function associated with the mth answer option for item i. We see in this formulation two actions: (1) the exponential transform that ensures that the probability will be positive, and (2) the normalization by dividing by sum $exp(W)$ in order to ensure that the probability values sum to one. The optimal score is found by minimizing the value of θ as defined by the following equation

$$\frac{dH}{d\theta} = -\sum_{i=1}^{n}\left[\sum_{m-1}^{M_i}[U_{im} - P_{im}(\theta)]\frac{dW_{im}}{d\theta}\right] = 0, \cdot \quad (2)$$

where $U_{im} - P_{im}$ is the difference between the data and the model fit and

$$dW_{im}/d\theta$$

is a coefficient that gives more weight if the item option contributes more to the knowledge of the test taker's ability. For more computational details of full information optimal scores, please refer to Ramsay et al. (2020).

2.2 The SweSAT

To examine student performance and to predict their success in university we used scores from the college admissions test SweSAT. The test is optional and is given twice a year. The test taker can repeat the test as many times as they prefer as only the best results counts. The test results are valid for five years. The SweSAT contains 160 multiple choice items and is divided into one verbal section and one quantitative section with 80 items each. The verbal section contains Vocabulary (20 items), Swedish reading comprehension (20 items), English reading comprehension (20 items) and Sentence completion (20 items). The quantitative section contained Data sufficiency (12 items), Diagrams, tables and maps (24 items), Mathematical problem solving (24 items) and Quantitative comparisons (20 items). Sum scores are used to calculate the test takers score on the test. There are 4–5 response alternatives to each of the multiple-choice items.

2.3 Participants

We used samples of students who were admitted to four different higher education programs in Sweden. The programs were chosen as they have different variations in their test score distributions and the chosen programs were: biomedical analysts (biomed), civil engineering (cing), college engineering (hing), and medical program (medical). The distribution of students at the different examined programs are given in the result section in Table 1.

The following student background variables were examined. Migration defined as 1 if the student or at least one of the students' parents were born in Sweden and 0 otherwise. Boys were coded as 1 and girls were coded as 0. Educational background was coded as 0 if the student had high school education or lower, and it was coded as 1 if the student had any post high school education. As it is also possible to get admitted to a university program in Sweden using only high school

grades, we used a grade variable which was composed by the final high school grade for each student. The grade variable was a constructed grade variable from the years 1997–2012, as we had high school graduates from all those years. The constructed grade variable was used so that all students were placed on the same grading scale even though the grade scale has changed in Sweden during those years. A grade A is equivalent to 20, a grade B is equivalent to 17.5, a grade C is equivalent to 15, a grade D is equivalent to 12.5, grade E is equivalent to 10 and the grade F is equivalent to 0. A student can also get extra credits (0.40) for extra curriculum activities. This means that the grade point average has a range of 0.0–20.40.

To get a measure of the students' achievement on their college education program we used a constructed variable, Relprest which have been used in other validation studies (e.g. Lyrén et al., 2014). Relprest is defined as the ratio between the students' passed credits and registered credits in their first year of college or university. The range of Relprest is 0.0–2.0, as students get zero if they do not take any of the credits they signed up for and some students have signed up for twice as many credits as the normal study rate.

2.4 Statistical Analysis

In the analyses we used both SweSAT sum scores and SweSAT optimal scores. We started by examining the score distributions using histograms and to examine the linear relationship between the two test scores we used scatterplots. Next, we examined the linear relationship between the test scores and Relprest with Pearson correlation. To examine the possible predictive effect, we used linear regressions with *Relprest* as dependent variable and the different test scores together with the students' background variables as independent variables. We also examined the test score distributions of the top 10% students with respect to their sum scores. The optimal scores were calculated using TestGardener (Ramsay & Li, 2021; Li et al., 2019) and the other statistical analyses were done in SPSS.

3 Results

Figure 1 displays the test score distributions for sum scores and optimal scores and Fig. 2 gives the scatterplot for the whole sample of those who took the SweSAT. From this figure it is clear that the distributions are not exactly the same. The distributions however share some similar features as the mean of the sum scores was 94.26 (SD = 21.98, Range: 32–151) and the mean for optimal scores was 94.39 (SD = 21.95, Range: 40.12–142.79). Although the sum scores have lower minimum and higher maximum than the optimal scores, the mid score range is a bit more flatten for the optimal score distribution as compared with the sum score distribution. The upper score range also differed depending on used test score.

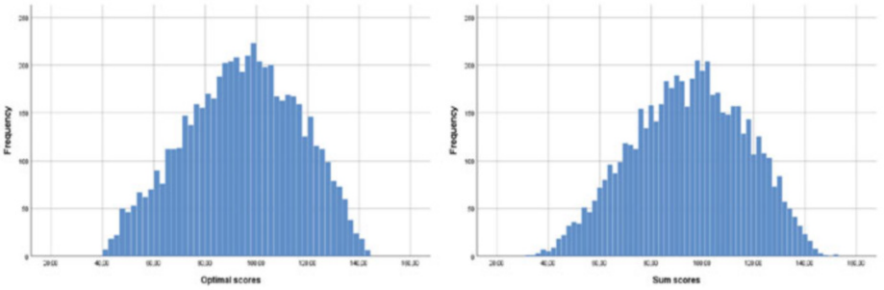

Fig. 1 Test score distributions with optimal scores to the left and sum scores to the right

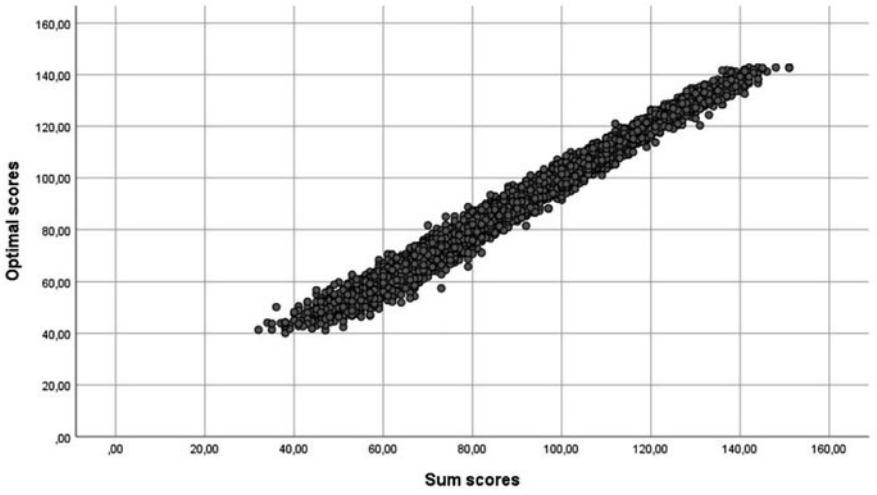

Fig. 2 Scatterplot between optimal scores and sum scores

The left part of Table 1 shows the correlation between full information optimal scores, sum scores and *Relprest* for the total SweSAT and the two SweSAT subsections; Quantitative and Verbal. From this table it is evident that there are overall quite small differences between optimal scores and sum scores. The non-significant correlations for the medical program are probably due to the fact that the variations of test scores are small in the medical programs. The right part of Table 1 gives the correlation between the students' high school grades and *Relprest*. Again, the weak correlation for medical students is due to the small variation of high school grades in this group.

To study the linear correlation between the SweSAT optimal score and the SweSAT sum scores we refer to Fig. 3 for the four university programs of interest. The correlations were very high and ranged between 0.98 (Biomed and Hing) to 0.99 (Cing and Medical). From these numbers and Fig. 3, it is evident that the sum scores and optimal scores are highly correlated but the scores differ for most of the

Table 1 Correlations between full information optimal scores, sum scores and *Relprest* in the left columns for the total SweSAT and the two subsections. The right columns shows the correlations between grades and *Relprest*

		Total		Quant		Verbal		Grade		
Exam	n	*r*	Sig.	r	Sig.	*r*	Sig.	n	r	Sig.
Biomed										
Optimal	178	0.39	***	0.30	***	0.38	***	149	0.44	***
Sum	178	0.38	***	0.28	***	0.38	***			***
Cing										
Optimal	3172	0.22	***	0.26	***	0.12	***	3036	0.38	***
Sum	3172	0.21	***	0.25	***	0.12	***			***
Hing										
Optimal	1404	0.20	***	0.22	***	0.13	***	1297	0.38	***
Sum	1404	0.19	***	0.21	***	0.13	***			***
Medical										
Optimal	827	−0.03	NS	0.00	NS	−0.05	NS	740	0.24	***
Sum	827	0.01	NS	0.01	NS	−0.05	NS			***

Biomed = Biomedical analytics, Cing = civil engineering, Hing = College engineering, Medical = Medical program. Total = Total SweSAT scores. Quant = Quantitative section scores, Verb = Verbal section scores. Grade = Correlation between Relprest and grades. NS = non-significant
*** = *p*-value less than 0.01

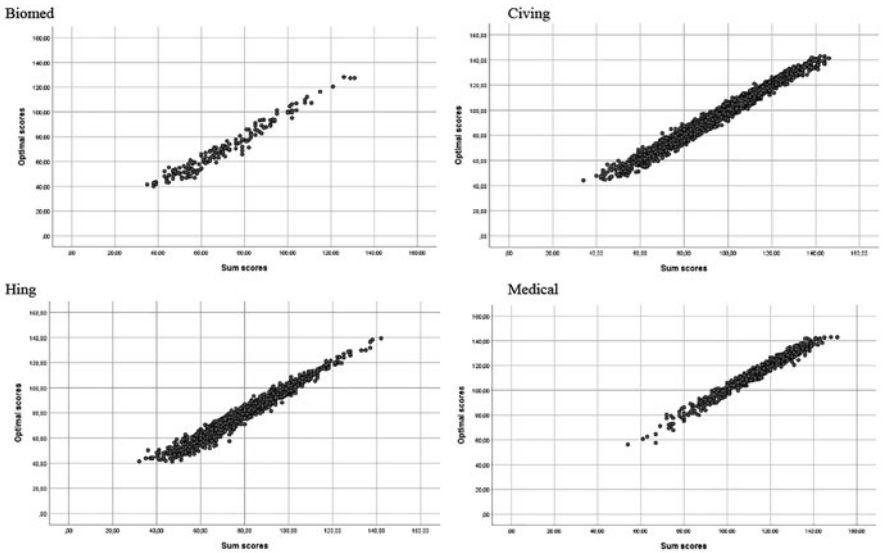

Fig. 3 The relationship between SweSAT optimal scores and SweSAT sum scores in the four different university programs

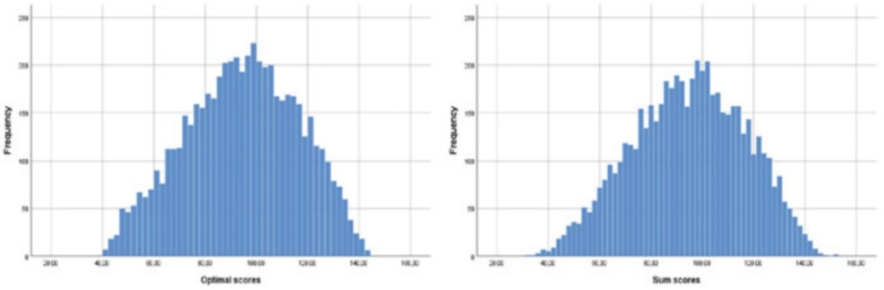

Fig. 4 Optimal test scores and sum score distributions for top 10% performers chosen from SweSAT sum scores

test takers. The differences differ over the score range depending on the university program and the score difference can be as large as 10 score points.

As SweSAT is primary used as a higher education admissions test, higher scores are of more interest than the lower score range. The 10% top performers admitted to the programs as defined from the sum scores are given in Fig. 4. From these plots it is evident that the top performers have slightly different test score distributions.

To further examine the predictive validity of the SweSAT scores we used the variable Relprest which is an indication of how the admitted students performed in their first year of university in comparison to what courses they have signed up for. We examined several student background variables, but the students' gender or educational background was never significant in any of the examined programs and thus the result is excluded from the table. The reason for the non-significance of the educational background is probably due to the rough definition of this variable as it only stated whether or not you have studied anything after high school or not. In Table 2, linear regressions with *Relprest* as dependent variable with optimal scores on every second line and sum scores on the other lines are given. We examined three different linear regression models. In model 1, only either optimal scores or sum scores were used as independent variable. In model 2, we used the test scores together with the grade variable. Finally, in model 3 we included the test scores, grades and the students' migration home background. The best fitting model was model 2 for all university programs and for the total SweSAT as well as for the subsections. Model 3 had many non-significant variables for most of the examined programs, regardless of test score used. We give the value of R from the linear regressions in the table and note that the values are very similar regardless of the test score used. Only small differences are shown and those were mainly when model 1 was used. Again, a reason for the non-significant values for the medical program is probably due to the very small variation in test scores for those who got admitted to the program.

Table 2 The R values and the sample sizes (n) from the three different linear regression models with *Relprest* as dependent variable with optimal scores on every second line and sum scores the other lines

		Model 1			Model 2 (G)			Model 3 (GI)		
Exam	n	Tot	Q	V	Tot	Q	V	Tot	Q	V
Biomed										
Optimal	178	0.39**	0.30**	0.38**	0.46**	0.44	0.47**	0.47	0.47	0.49
Sum	178	0.38**	0.28**	0.38**	0.46**	0.44	0.48**	0.47	0.47	0.49
Cing										
Optimal	3172	0.22**	0.26**	0.12**	0.39**	0.41**	0.39	0.39	0.41*	0.39
Sum	3172	0.21**	0.25**	0.12**	0.39**	0.41**	0.38	0.39	0.41*	0.39
Hing										
Optimal	1404	0.20**	0.22**	0.13**	0.39**	0.40**	0.38**	0.40**	0.41**	0.39
Sum	1404	0.20**	0.21**	0.13**	0.39**	0.40**	0.38**	0.40**	0.41**	0.39
Medical										
Optimal	827	0.03	0.00	0.05	0.24	0.24	0.24	0.25	0.25	0.25
Sum	827	0.02	0.01	0.05	0.24	0.24	0.24	0.25	0.25	0.25
All										
Optimal	5581	0.25**	0.26**	0.19**	0.40**	0.40**	0.40**	0.40**	0.41**	0.40**
Sum	5581	0.25**	0.26**	0.19**	0.40**	0.40**	0.39**	0.40**	0.41**	0.40**

Biomed = Biomedical analytics, Cing = civil engineering, Hing = College engineering, Medical = Medical program. N = Number of test takers. Tot = Total test score. Q = Quantitative test score, V = Verbal test score. Sum = Sum scores are used as independent variable instead of optimal scores. G = Grade, I = Immigration

** = p-value less than 0.01, * = p-value less than 0.1

4 Discussion

The overall aim was to examine the predictive validity and thus the relationship between students' college admissions test results and the students' performance in higher education using sum scores and optimal scores. The results indicated that both optimal scores and sum scores can predict the students' university performance similarly regardless if we control for some covariates or not. Although the test score distributions differed, the overall results of predictivity of the students' performance were similar. This is good news as it means that optimal scores can be used in these situations. Although the overall conclusions were similar, the test score distributions differed in the sense that the optimal score distribution had a slightly more flattened curve than the sum score distribution. This means that for a certain student it may have impact which test score is used when the test score is used for selection to higher education even though a clear difference is not seen on the overall results. The differences between test results based on sum scores and optimal scores are typically larger for programs which require high test scores. The result that different test takers may get different sum scores and optimal scores are inline with previous

studies of optimal scores (Ramsay & Wiberg, 2017a,b; Wiberg et al., 2018). The impact for a specific student should be addressed further in the future.

There were a few limitations of this study. First, as a measure of success in higher education we used a relative performance measure of the students' performance. This measure is probably a bit blunt and thus future studies should probably use a more refined measurement. However, this measure was used in the SweSAT prognosis study by Lyrén et al. (2014), which obtained similar result for the sum scores as in our study. Note, some numbers concerning the sum scores differed between our study and their study as we in contrast to them only used complete cases. Second, we only had access to a few student background variables and in future studies, it would be of interest to include other background variables. Third, in this study we only had access to those admitted to the university programs and thus the study has a range restriction. As the optimal scores and sum scores differ in their distributions, it is likely that the rank of the students differ within the test scores. If one would change from sum scores to optimal scores it is likely that some students may have not admitted to a program and others were admitted to a program and thus the choice of test score could potential influence the life of a student. However, on a group level the results are similar and thus one should be comfortable to use either sum scores or optimal scores in admissions tests. An advantage of using optimal scores in sum scores, seen in e.g. Ramsay and Wiberg (2017b), is that the precision of estimating the ability of the students is better and thus optimal scores should be considered for high stakes test as it would be a fairer instrument for the students.

Acknowledgement The research was funded by the Swedish Wallenberg MMW 2019.0129 grant.

References

Cliffordson, C. (2008). Differential prediction of study success across academic programs in the swedish context: The validity of grades and tests as selection instruments for higher education. *Educational Assessment, 13*(1), 56–75.

Cliffordson, C., & Askling, B. (2006). Different grounds for admission: Its effects on recruitment and achievement in medical education. *Scandinavian Journal of Educational Research, 50*(1), 45–62.

Dorans, N. J. (1999). Correspondences between act and sat®i scores. *ETS Research Report Series, 1999*(1), i–18.

GRE (2021). *Guide to the Graduate Record Exams (GRE)*. https://www.ets.org/s/gre/pdf/gre_guide.pdf. Accessed 19 Jan 2021.

Li, J., Ramsay, J. O., & Wiberg, M. (2019). TestGardener: A program for optimal scoring and graphical analysis. In *Quantitative Psychology 83rd Annual Meeting of the Psychometric Society* (pp. 87–94). Springer.

Lyrén, P.-E. (2008). Prediction of academic performance by means of the swedish scholastic assessment test. *Scandinavian Journal of Educational Research, 52*(6), 565–581.

Lyrén, P.-E., Rolfsman, E., Wedman, J., Wikström, C., & Wikström, M. (2014). *Det nya högskoleprovet: samband mellan provresultat och prestation i högskolan* [*The new SweSAT: Association between test results and performance in higher education*]. Umeå, Sweden: Department of Applied Educational Science, Umeå University.

Ramsay, J. O., & Li, J. (2021). TestGardener: Optimal Analysis of Test and Rating Scale Data. R package version 2.0.1 https://CRAN.R-project.org/package=TestGardener

Ramsay, J. O., & Wiberg, M. (2017a). Breaking through the sum scoring barrier. In *Quantitative Psychology 81st Annual Meeting of the Psychometric Society* (pp. 151–158). Springer.

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

Ramsay, J. O., Wiberg, M., & Li, J. (2020). Full information optimal scoring. *Journal of Educational and Behavioral Statistics, 45*(3), 297–315.

Svensson, A., Gustafsson, J.-E. and Reuterberg, S.-E. (2001). *Högskoleprovets prognosvärde. Samband mellan provresultat och framgång första året vid civilingenjörs-, jurist- och grundskollärarutbildningarna* [*The prognostic value of the SweSAT. Association between test result and first-year academic performance at the civil engineering, law and elementary school teacher programmes*] (Högskoleverkets rapportserie No. 2001:19 R). Stockholm: Högskoleverket.

Wiberg, M., Ramsay, J. O., & Li, J. (2018). Optimal scores as an alternative to sum scores. In *Quantitative Psychology 82nd Annual Meeting of the Psychometric Society* (pp. 1–10). Springer.

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.