PLOS ONE

RESEARCH ARTICLE

Evaluating the cost of simplicity in score building: An example from alcohol research

Valentin Rousson[1], Bastien Trächsel[1], Katia Iglesias[2], Stéphanie Baggio[3,4]*

1 Division of Biostatistics, Center for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland, **2** School of Health Sciences Fribourg (HEdS-FR), HES-SO University of Applied Sciences and Arts of Western Switzerland, Fribourg, Switzerland, **3** Institute of Primary Health Care (BIHAM), University of Bern, Bern, Switzerland, **4** Laboratory of Population Health (#PopHealthLab), University of Fribourg, Fribourg, Switzerland

* stephanie.baggio@unibe.ch

Citation: Rousson V, Trächsel B, Iglesias K, Baggio S (2023) Evaluating the cost of simplicity in score building: An example from alcohol research. PLoS ONE 18(11): e0294671. https://doi.org/10.1371/journal.pone.0294671

Editor: Sathishkumar Veerappampalayam Easwaramoorthy, Sunway University, MALAYSIA

Received: March 21, 2023

Accepted: October 20, 2023

Published: November 27, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0294671

Data Availability Statement: Data and code are available as supplementary materials.

Funding: This study was supported by Swiss National Research Foundation (no. 10001C_173418/1). The funders had no role in

Abstract

Building a score from a questionnaire to predict a binary gold standard is a common research question in psychology and health sciences. When building this score, researchers may have to choose between statistical performance and simplicity. A practical question is to what extent it is worth sacrificing the former to improve the latter. We investigated this research question using real data, in which the aim was to predict an alcohol use disorder (AUD) diagnosis from 20 self-reported binary questions in young Swiss men (n = 233, mean age = 26). We compared the statistical performance using the area under the ROC curve (AUC) of (a) a "refined score" obtained by logistic regression and several simplified versions of it ("simple scores"): with (b) 3, (c) 2, and (d) 1 digit(s), and (e) a "sum score" that did not allow negative coefficients. We used four estimation methods: (a) maximum likelihood, (b) backward selection, (c) LASSO, and (d) ridge penalty. We also used bootstrap procedures to correct for optimism. Simple scores, especially sum scores, performed almost identically or even slightly better than the refined score (respective ranges of corrected AUCs for refined and sum scores: 0.828–0.848, 0.835–0.850), with the best performance been achieved by LASSO. Our example data demonstrated that simplifying a score to predict a binary outcome does not necessarily imply a major loss in statistical performance, while it may improve its implementation, interpretation, and acceptability. Our study thus provides further empirical evidence of the potential benefits of using sum scores in psychology and health sciences.

Introduction

Composite scores are widely used in psychology and health sciences. Guidelines are available for the development and validation of these scores, but recommendations for analytical strategies are less common [1]. Composite scores can be calculated at different levels of complexity [2]. The simplest composite score would be a sum score, in which the possible values are restricted to be either +1 or 0. In this sum score, all questions with a non-zero coefficient have the same positive weight. More sophisticated approaches to composite scores include the use

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

of restricted value ranges (e.g., +1, -1 or 0) or linear combinations of the items. Such composite scores ("refined scores") can be developed using logistic regression models with a gold standard as the response variable and the items as predictors. These approaches allow for unequal weighting of questions.

Controversy on simple scores

The use of simple or refined scores has been much discussed and is still currently debated. First, it has been discussed in the context of scores obtained from factor or principal component analyses [3–7] with conflicting conclusions. While two recent studies warned that sum scores may be too imprecise for use in rigorous research applications [8, 9], another study presented an example where little was gained from the use of factor score estimates (i.e., refined scores) compared to simpler sum scores [10]. A third opinion paper also concluded that sum scores are suitable to build scores [2]. In the context of linear regression modeling, previous studies suggested that equal regression weights might be a reasonable choice [11–13], especially if predictors are standardized, with a modest loss of accuracy compared to unequal weight [11]. To our knowledge, the use of simple or refined scores was not discussed in the context of logistic regression. Further empirical investigations are therefore needed to better understand the benefits and limitations of simple or refined scores in this analytical context, as stated in recent studies [10, 13].

Understudied perspectives

An interesting perspective that has been neglected in previous research is to identify the *cost of simplicity*. To facilitate implementation, interpretation, and acceptability, simple scores sacrifice some of the statistical performance for the sake of simplicity. If the loss of statistical performance does not appear to be substantial, this would argue in favor of using a simple rather than a refined score.

In addition, when evaluating the statistical performance of a score, it is important to consider problems of overfitting, also known as optimism, and to attempt to correct for them [14]. Overfitting occurs, for example, when a regression model includes too many predictors, but also when it is selected from a large family of candidate models, e.g., via automated variable selection [15]. Overfitting may lead to replicability issues, a critical issue in psychology and health sciences [16]. Simple scores may be less prone to overfitting than refined scores. This may be an unexplored advantage of simple scores over refined scores.

Objective of the study

The aim of the present study was therefore to investigate and evaluate the cost of simplicity using real-life data, where the aim was to predict a diagnosis of alcohol use disorder (AUD) diagnosis from 20 self-reported binary questions. Unlike factor analysis, where a score is developed to measure a theoretical construct that is not observable, we had the advantage of having a gold standard against which to compare our predictions. It was therefore possible to objectively compare the statistical performance (including optimism) of refined and simple scores obtained by different methods.

Materials and methods

Design

We re-used data from a prospective cross-sectional study designed to identify an accurate screening tool for AUD [17, 18]. The study was approved by the Ethics Committee of the

Canton of Vaud (no. 2017–00776). Participants signed a written informed consent for the study and an additional consent form to accept the reuse of their data in further projects. The authors did not have access to any information that could identify individual participants during or after data collection.

Participants

Data were collected from October 2017 to June 2018 in a sample of young Swiss men. They were recruited from the Cohort Study on Substance Use and Risk Factors (C-SURF) [19]. Inclusion criteria were 1) being a French-speaking participant, 2) completing the second follow-up questionnaire (from 2016 to 2018), and 3) having a valid email (n = 2,668). Eligible participants were invited to complete the ten-question version of the Alcohol Use Disorder Identification Test (AUDIT) [20] online (1,371 respondents, response rate = 51.4%). Participants were then selected using a stratified sampling strategy: those with a high AUDIT score ($\geq$13) and those with a low score ($<$13) [21]. The final sample size was 233 (total response rate = 70.6%, 68.9% in the low-strata group and 72.0% in the high-strata group).

Diagnosis of AUD

A binary variable measured the presence or absence of AUD, assessed with a clinician-administered diagnostic interview (Diagnostic Interview for Genetic Studies (DIGS) [22]) and representing the gold standard. The DIGS has a high inter-rater agreement and a good concordance with clinical diagnoses from medical records [22]. At the time of the study, the DIGS had not been adapted to the DSM-5 criteria. To address this limitation, we replaced the DSM-IV question on legal problems (removed in DSM-5) with a question on craving (added in DSM-5). AUD was defined as at least mild (cut-off score = 2) in the previous twelve months.

Self-reported AUD and alcohol-related consequences

A set of 20 binary questions (1 = yes/0 = no, hereafter Q1-Q20) designed to screen for AUD was used to predict the gold standard. Participants self-reported the presence or absence of the eleven DSM-5 AUD criteria [20, 23] and of nine alcohol-related consequences [20, 24, 25] in the previous twelve months. The questions are listed in S1 Table.

Analytical strategy

The sample size was calculated for the original study purpose [17, 18]. As AUD was overrepresented in our study sample, we focused on discrimination rather than on calibration [14] when assessing the statistical performance of our scores. Score performance was measured using the area under the ROC curve (AUC) [26].

Refined score

Our aim was to build a score that best predicted the gold standard (AUD) from the responses given to questions Q1-Q20. We fitted a logistic regression model with the gold standard as the outcome and questions Q1-Q20 as binary predictors. Coefficients were used for the score.

Simple scores

We defined four simple scores. First, we simplified non-zero coefficients with m = 3, 2 or 1 possible digit(s) (see S1 File for details). A fourth simple score allowed zero or positive coefficients, but not negative coefficients, is called a "sum score". This is because in our example data, all 20 predictors were designed to be positively associated with the gold standard. In such

a context, having negative coefficients may undermine the acceptability of a simple score, so it is tempting to remove negative coefficients (set them to zero). This is consistent with the recommendation of Steyerberg et al. [27], who advocate "using qualitative information on the sign of the effect of predictors".

Methods

For both refined and simple scores, we first used maximum likelihood estimation (MLE). Then, to reduce the number of predictors in the model, we used the well-known backward elimination procedure (hereafter BACKWARD), which consists of starting with a model including all the predictors and eliminating the least significant predictor at each step of an iterative procedure. We used the Akaike criterion to select the best model [28]. We also used other more modern methods for fitting a model with many predictors with penalized maximum likelihood, i.e., the LASSO or RIDGE penalty (also called the L1 or L2 penalty, respectively), where the coefficients defining the score are shrunk towards zero [29]. The LASSO penalty sets some coefficients exactly to zero, making the resulting score more parsimonious.

Finally, the results may be too optimistic. One reason is that our scores were derived and evaluated from the same data. For BACKWARD, another reason is that we used a strict model selection. To correct for optimism, we applied a bootstrapping procedure, as described and recommended by Steyerberg et al. [30]. Note that for BACKWARD and LASSO, the resulting model did not necessarily include the same number of predictors in each bootstrap resample.

In each resample and for each method, we calculated two AUCs: one using the data from the bootstrap resample and one using the data from the original sample. Optimism was estimated as the difference between these two AUCs, averaged over the 500 bootstrap resamples.

Analyses were performed using R software. For the LASSO and RIDGE procedures, we applied the default parameters implemented in the *glmnet* library (version 4.1–3). The statistical code and dataset are available as supplementary material.

Results

The proportion of patients with AUD was 33.5% (n = 78) (mean age = 27.00). The proportion of patients answering "yes" to the different questions ranged from 4% (Q18) to 67% (Q6). All questions were significantly positively associated with the gold standard, with odds-ratios ranging from 1.89 (for Q20) to 13.08 (for Q11), except for one question (Q20). Table 1 summarizes this information and also shows the sensitivity and specificity achieved by each question, as well as the AUC, which ranged from 0.525 (for Q20) to 0.679 (for Q4).

The main results for the refined and simples scores and different methods are shown in Table 2. For the refined score, the AUC for MLE was 0.890, higher than the AUCs when each individual question was considered as a predictor. Using the BACKWARD procedure, the final model included eight questions (Q2, Q4, Q5, Q8, Q9, Q10, Q14 and Q15) and the AUC was 0.876. We obtained AUCs of 0.887 for LASSO (Q1, Q2, Q4, Q5, Q7, Q8, Q10, Q11, Q12, Q13, Q14, Q15 and Q17, other coefficients set to zero) and 0.881 for RIDGE. The coefficients assigned to the different questions for the four methods considered (MLE, BACKWARD, LASSO, RIDGE) are plotted in the four panels of the first column of Fig 1. None of these scores are simple since all non-zero coefficients are different from each other. It is worth noting that some coefficients were negative for MLE and RIDGE.

In bootstrap analyses, we estimated an optimism of 0.052, 0.047, 0.045 and 0.039 for MLE, BACKWARD, LASSSO and RIDGE, respectively. Finally, corrected AUCs were obtained by subtracting the estimated optimism from the observed AUCs, yielding 0.838, 0.828, 0.848 and 0.846, respectively (see Table 2).

Table 1. Summary of the associations between the 20 questions and the gold standard.

Question	Proportion yes	Sensitivity	Specificity	AUC	OR	P-value
Q1	75/233 = 32%	45/78 = 58%	125/155 = 81%	0.692	5.68	< .001
Q2	150/233 = 64%	64/78 = 82%	69/155 = 45%	0.633	3.67	< .001
Q3	27/233 = 12%	20/78 = 26%	148/155 = 95%	0.606	7.29	< .001
Q4	49/233 = 21%	35/78 = 45%	141/155 = 91%	0.679	8.20	< .001
Q5	27/233 = 12%	21/78 = 27%	149/155 = 96%	0.615	9.15	< .001
Q6	155/233 = 67%	64/78 = 82%	64/155 = 41%	0.617	3.22	.001
Q7	24/233 = 10%	16/78 = 21%	147/155 = 95%	0.577	4.74	.001
Q8	77/233 = 33%	46/78 = 59%	124/155 = 80%	0.695	5.75	< .001
Q9	12/233 = 5%	9/78 = 12%	152/155 = 98%	0.548	6.61	.006
Q10	30/233 = 13%	23/78 = 29%	148/155 = 95%	0.625	8.84	< .001
Q11	19/233 = 8%	16/78 = 21%	152/155 = 98%	0.593	13.08	< .001
Q12	55/233 = 24%	31/78 = 40%	131/155 = 85%	0.621	3.60	< .001
Q13	110/233 = 47%	54/78 = 69/	99/155 = 64%	0.666	3.98	< .001
Q14	108/233 = 46%	54/78 = 69%	101/155 = 65%	0.672	4.21	< .001
Q15	56/233 = 24%	33/78 = 42%	132/155 = 85%	0.637	4.21	< .001
Q16	28/233 = 20%	15/78 = 19%	142/155 = 92%	0.554	2.60	.019
Q17	46/233 = 12%	28/78 = 36%	137/155 = 88%	0.621	4.26	< .001
Q18	10/233 = 4%	7/78 = 9%	152/155 = 98%	0.535	5.00	.022
Q19	37/233 = 16%	22/78 = 28%	140/155 = 90%	0.593	3.67	< .001
Q20	19/233 = 8%	9/78 = 12%	145/155 = 94%	0.525	1.89	.186

AUC: Area under the curve; OR: odds-ratio.

https://doi.org/10.1371/journal.pone.0294671.t001

The simple scores obtained with m = 3, 2 or 1 possible digit(s) are plotted in the second, third, and fourth columns of Fig 1 for the four methods. Unlike the refined score plotted in the first column of Fig 1, the non-zero coefficients of these simple scores are not all different from each other.

Observed and corrected AUCs for all simple scores are shown in Table 2. Optimism was systematically lower for simple scores than for the refined score. Among the simple scores, the optimism was also systematically lower with m = 1 than with m = 2 digits and with m = 2 than

Table 2. Observed / corrected AUCs for refined and simple scores to predict the gold standard using various methods.

		Scores				
		refined	**simple-3d**	**simple-2d**	**simple-1d**	**sum**
Methods	MLE	0.890 / 0.838	0.881 / 0.834	0.878 / 0.833	0.856 / 0.816	0.864 / 0.848
		(-0.052)	(-0.047)	(-0.045)	(-0.039)	(-0.016)
	BACKWARD	0.876 / 0.828	0.873 / 0.828	0.873 / 0.829	0.866 / 0.826	0.866 / 0.835
		(-0.047)	(-0.045)	(-0.044)	(-0.040)	(-0.031)
	LASSO	0.887 / 0.848	0.886 / 0.852	0.881 / 0.850	0.870 / 0.843	0.870 / 0.849
		(-0.038)	(-0.034)	(-0.031)	(-0.027)	(-0.021)
	RIDGE	0.881 / 0.846	0.877 / 0.845	0.874 / 0.844	0.862 / 0.837	0.862 / 0.850
		(-0.035)	(-0.032)	(-0.030)	(-0.025)	(-0.012)

The optimism, which is the difference between the observed and the corrected AUC (obtained using a bootstrap procedure), is given in parentheses.

AUC: Area under the curve; MLE: maximum likelihood estimation; BACKWARD: backward elimination procedure, selection using Akaike criterion; LASSO: L1-penalized maximum likelihood; RIDGE: L2-penalized maximum likelihood.

Refined score: coefficients from logistic regression; simple-3d, 2d, and 1d scores: simple scores with 3, 2, and 1 digit(s); sum score: negative coefficients set to zero and 1 digit.

https://doi.org/10.1371/journal.pone.0294671.t002

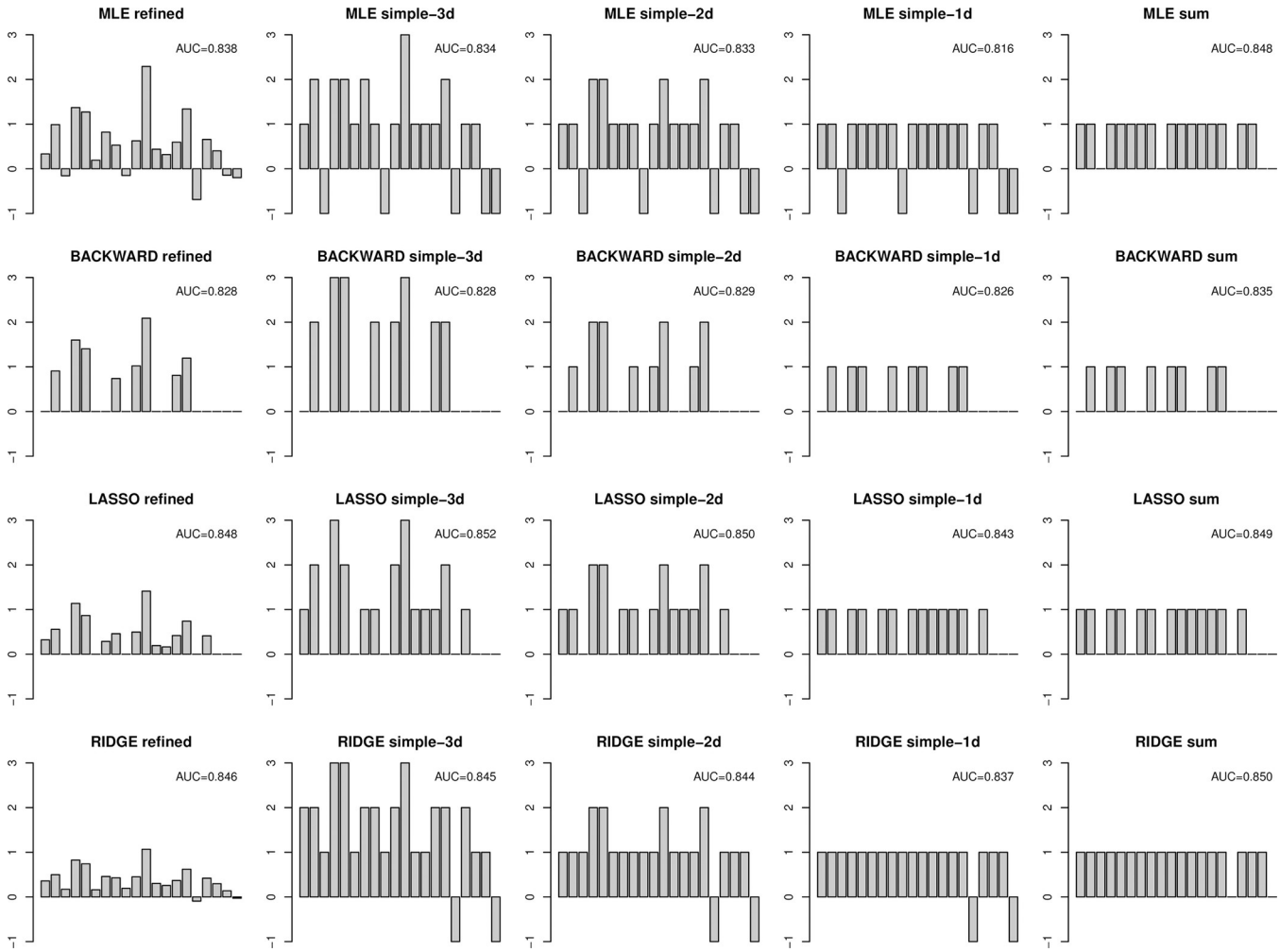

Fig 1. Graphical representation of the coefficients of refined and simple scores to predict the gold standard obtained using various methods, together with corrected AUCs. In each panel, the 20 bars represent the coefficients assigned to questions Q1-Q20 (from left to right). AUC: Area under the curve; MLE: maximum likelihood estimation; BACKWARD: backward elimination procedure, selection using Akaike criterion; LASSO: L1-penalized maximum likelihood; RIDGE: L2-penalized maximum likelihood. Refined score: coefficients from logistic regression; simple-3d, 2d, and 1d scores: simple scores with 3, 2, and 1 digit(s); sum score: negative coefficients set to zero and 1 digit.

https://doi.org/10.1371/journal.pone.0294671.g001

with m = 3 digits. After correcting for optimism, the best performance in terms of AUC was obtained with the simplified LASSO with m = 3 digits, with a corrected AUC of 0.852, which was even better than the refined score obtained with LASSO (with a corrected AUC of 0.848).

Finally, the fifth column of Fig 1 shows the sum scores obtained by the four methods by setting the negative coefficients to zero. For BACKWARD and LASSO, they were identical to the one-digit simple scores. These scores are sums of 15, 8, 13 and 18 questions for MLE, BACKWARD, LASSO and RIDGE, respectively. The observed and corrected AUCs for these sum scores are shown in Table 2. Optimism was even lower for sum scores than for simple scores. The corrected AUCs for sum scores were of 0.848, 0.835, 0.849 and 0.850, respectively. Except for BACKWARD, which was slightly above, the sum scores obtained via MLE, LASSO or RIDGE performed almost as well (or even better than) the refined score via LASSO (with a corrected AUC of 0.848). In particular, the sum score via LASSO with a corrected AUC of 0.849 and only 13 questions could be a good final choice for this example, as the resulting score is not only simple but also parsimonious.

It should be noted that the sum score used by Baggio et al. [17] for these data included 12 instead of 13 questions. It was not obtained as a simplified version of a refined score, but as the sum score minimizing the Akaike criterion among all 2^20–1 = 1'048'575 possible sums, achieving an observed AUC of 0.872 and a corrected AUC of 0.841.

Discussion

In this study, we attempted to simplify and evaluate the statistical performance in terms of AUC of refined and simple scores obtained by different methods using data from alcohol research where the aim was to predict an AUD from 20 binary questions.

Among the refined score methods, the best performance was achieved by LASSO with a corrected AUC of 0. 848. The MLE method had the highest observed AUC (0.890), but it was the most optimistic method (i.e., the most prone to overfitting). However, as we only had 78 cases of AUD in our dataset, a model with 20 predictors did not follow the rule of thumb of 10 required events per predictor. This could lead to overfitting, although this rule of thumb should not be taken too strictly and has recently been questioned [31].

Among the ways of defining simple scores and related methods, a simple score with 3 digits using LASSO was the best, even better than the refined score (corrected AUC = 0.852). Other simple scores, especially sum scores, performed almost identically or even slightly better than the refined score obtained with some methods, illustrating the fact that simplifying a score does not necessarily imply a major loss in statistical performance. Indeed, the sum score had the highest corrected AUCs for MLE, BACKWARD, and RIDGE.

Overall, the more constrained the coefficients were, the less prone a method was to overfitting. Simple scores, and even more, sum scores, are less prone to overfitting than the refined score because they are less data-dependent due to the restrictions imposed on their possible values. Therefore, the simplification of a refined score does not necessarily come at the cost of sacrificing statistical performance. Such considerations were anticipated by in a previous study in the context of linear regression [11] and factor analysis [10]. The latter found that factor scores could lead to greater indeterminacy than sum scores [10]. Estimates of the former may vary from sample to sample, whereas sum scores have identical weights in all samples. Our study illustrated and confirmed this finding in the context of logistic regression.

It should be noted that the corrected AUCs provided in our data are only sample estimates of the true AUC that would be obtained by applying a score to the entire population of interest. As the confidence intervals calculated around them would largely overlap, we would not be able to conclude that the corrected AUCs for a simple score would be "significantly higher" than the corrected AUCs for the refined score. However, the reverse would also be true (the corrected AUCs for the refined score would not be significantly higher than the corrected AUCs for some simple scores) and this may already be sufficient justification for using a simple score in practice. In any case, researchers should not be discouraged a priori from striving for simplicity, as the sacrifice of statistical performance may be very small. This is in line with recent conclusions [10], which also noted that sum scores are easier to implement than factor scores.

These recommendations are based on a single, albeit real, data example, which constitutes a study limitation. In addition, the sample size was relatively small, which did not allow splitting the dataset into train and test sets. Therefore, we do not claim or advocated that it is possible to replace refined scores with simple scores in all practical cases. Rather, we encourage researchers and methodologists developing screening tools to evaluate the cost of simplicity along the lines presented here, including a correction for optimism, with their own data. Their results would provide a sound basis and some justification for deciding whether to retain a refined score or replace it with a simple score. Although researchers and statisticians have

different views on the use of simple scores, there is a consensus on the need to take psychometrics seriously and to provide justification for the preferred scoring methods [2, 8, 10, 16].

Conclusion

To conclude, our example data demonstrated that simplifying a score to predict a binary outcome does not necessarily imply a major loss in statistical performance, while potentially improving its implementation, interpretation, and acceptability. Our study thus provided further empirical evidence of the potential benefits of using sum scores in psychology and health sciences. Future studies should examine other practical or simulated cases to further evaluate the cost of simplicity and provide robust empirical evidence on this controversial issue.

Supporting information

S1 Checklist. *PLOS ONE* clinical studies checklist.
(DOCX)

S2 Checklist. STROBE statement—checklist of items that should be included in reports of observational studies.
(DOCX)

S1 Table. Items for self-reported alcohol use disorders and alcohol-related consequences.
(DOCX)

S1 File. Simplification algorithm.
(DOCX)

S2 File. R code.
(R)

S1 Data. Database.
(TXT)

Author Contributions

Conceptualization: Valentin Rousson, Bastien Trächsel, Katia Iglesias, Stéphanie Baggio.

Data curation: Katia Iglesias, Stéphanie Baggio.

Formal analysis: Valentin Rousson, Bastien Trächsel.

Funding acquisition: Stéphanie Baggio.

Investigation: Katia Iglesias.

Methodology: Valentin Rousson, Bastien Trächsel, Katia Iglesias, Stéphanie Baggio.

Project administration: Stéphanie Baggio.

Supervision: Valentin Rousson, Katia Iglesias.

Writing – original draft: Valentin Rousson.

Writing – review & editing: Bastien Trächsel, Katia Iglesias, Stéphanie Baggio.

References

1. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: A practical guide. Cambridge: Cambridge University Press; 2011 2011.

2. Edelsbrunner PA. A model and its fit lie in the eye of the beholder: Long live the sum score. Frontiers in Psychology. 2022;13. https://doi.org/10.3389/fpsyg.2022.986767 PMID: 36312188
3. DiStefano C, Zhu M, Mindrila D. Understanding and using factor scores: Considerations for the applied researcher. Practical Assessment, Research & Evaluation. 2009; 14:20.
4. Grice JW, Harris RJ. A comparison of regression and loading weights for the computation of factor scores. Multivariate Behavioral Research. 1998; 33(2):221–47. https://doi.org/10.1207/s15327906mbr3302_2 PMID: 26771884
5. Rousson V, Gasser T. Simple component analysis. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2004; 53(4):539–55. https://doi.org/10.1111/j.1467-9876.2004.05359.x
6. Vines SK. Simple principal components. Journal of the Royal Statistical Society Series C (Applied Statistics). 2000; 49(4):441–51.
7. Wackwitz JH, Horn JL. On obtaining the best estimates of factor scores within an ideal simple structure. Multivariate Behavioral Research. 1971; 6(4):389–408. https://doi.org/10.1207/s15327906mbr0604_2 PMID: 26825238
8. McNeish D, Wolf MG. Thinking twice about sum scores. Behavior Research Methods. 2020; 52 (6):2287–305. https://doi.org/10.3758/s13428-020-01398-0 PMID: 32323277
9. McNeish D. Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. Behavior Research Methods. 2022. https://doi.org/10.3758/s13428-022-02016-x PMID: 36394821
10. Widaman KF, Revelle W. Thinking thrice about sum scores, and then some more about measurement and analysis. Behavior Research Methods. 2022. https://doi.org/10.3758/s13428-022-01849-w PMID: 35469086
11. Wainer H. Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin. 1976; 83:213–7. https://doi.org/10.1037/0033-2909.83.2.213
12. Dawes RM. The robust beauty of improper linear models in decision making. American Psychologist. 1979; 34:571–82. https://doi.org/10.1037/0003-066X.34.7.571
13. Bobko P, Roth PL, Buster MA. The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. Organizational Research Methods. 2007; 10 (4):689–709. https://doi.org/10.1177/1094428106294734
14. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology. 2010; 21 (1):128–38. https://doi.org/10.1097/EDE.0b013e3181c30fb2 PMID: 20010215
15. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004; 66(3):411–21. https://doi.org/10.1097/01.psy.0000127692.23278.a9 PMID: 15184705
16. McNeish D. Limitations of the sum-and-alpha approach to measurement in behavioral research. Policy Insights from the Behavioral and Brain Sciences. 2022; 9(2):196–203. https://doi.org/10.1177/23727322221117144
17. Baggio S, Trächsel B, Rousson V, Rothen S, Studer J, Marmet S, et al. Identifying an accurate self-reported screening tool for alcohol use disorder: evidence from a Swiss, male population-based assessment. Addiction (Abingdon, England). 2020; 115(3):426–36. https://doi.org/10.1111/add.14864 PMID: 31656049
18. Iglesias K, Sporkert F, Daeppen J-B, Gmel G, Baggio S. Comparison of self-reported measures of alcohol-related dependence among young Swiss men: a study protocol for a cross-sectional controlled sample. BMJ Open. 2018; 8(7):e023632. https://doi.org/10.1136/bmjopen-2018-023632 PMID: 30012797
19. Gmel G, Akre C, Astudillo M, Bähler C, Baggio S, Bertholet N, et al. The Swiss Cohort Study on Substance Use Risk Factors–findings of two waves. SUCHT. 2015; 61(4):251–62. https://doi.org/10.1024/0939-5911.a000380
20. Knight JR, Wechsler H, Kuo M, Seibring M, Weitzman ER, Schuckit MA. Alcohol abuse and dependence among U.S. college students. J Stud Alcohol. 2002; 63(3):263–70. https://doi.org/10.15288/jsa.2002.63.263 PMID: 12086126
21. Meneses-Gaya C, Zuardi AW, Loureiro SR, Hallak JEC, Trzesniak C, de Azevedo Marques JM, et al. Is the full version of the AUDIT really necessary? Study of the validity and internal construct of its abbreviated versions. Alcohol Clin Exp Res. 2010; 34(8):1417–24. https://doi.org/10.1111/j.1530-0277.2010.01225.x PMID: 20491736
22. Berney A, Preisig M, Matthey M-L, Ferrero F, Fenton BT. Diagnostic interview for genetic studies (DIGS): inter-rater and test-retest reliability of alcohol and drug diagnoses. Drug and Alcohol Dependence. 2002; 65(2):149–58. https://doi.org/10.1016/s0376-8716(01)00156-9 PMID: 11772476

23. Grant BF, Dawson DA, Stinson FS, Chou PS, Kay W, Pickering R. The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample. Drug and Alcohol Dependence. 2003; 71(1):7–16. https://doi.org/10.1016/s0376-8716(03)00070-x PMID: 12821201

24. Dupuis M, Baggio S, Henchoz Y, Deline S, N'Goran A, Studer J, et al. Risky single occasion drinking frequency and alcohol-related consequences: can abstinence during early adulthood lead to alcohol problems? Swiss Med Wkly. 2014; 144:w14017. https://doi.org/10.4414/smw.2014.14017 PMID: 25295759

25. Wechsler H, Davenport A, Dowdall G, Moeykens B, Castillo S. Health and behavioral consequences of binge drinking in college. A national survey of students at 140 campuses. JAMA. 1994; 272(21):1672–7. PMID: 7966895

26. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biom J. 2008; 50(4):457–79. https://doi.org/10.1002/bimj.200810443 PMID: 18663757

27. Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making. 2001; 21(1):45–56. https://doi.org/10.1177/0272989X0102100106 PMID: 11206946

28. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G, editors. Selected Papers of Hirotugu Akaike. Springer Series in Statistics. New York, NY: Springer; 1998. p. 199–213.

29. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with applications in R: Springer; 2014 2014.

30. Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001; 54(8):774–81. https://doi.org/10.1016/s0895-4356(01)00341-9 PMID: 11470385

31. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Medical Research Methodology. 2016; 16(1):163. https://doi.org/10.1186/s12874-016-0267-3 PMID: 27881078