Scientometrics (2017) 111:581–593
DOI 10.1007/s11192-017-2283-4

How to reduce the number of rating scale items without predictability loss?

W. W. Koczkodaj[1] · T. Kakiashvili[2] · A. Szymańska[3] · J. Montero-Marin[4] · R. Araya[5] · J. Garcia-Campayo[6] · K. Rutkowski[7] · D. Strzałka[8]

Received: 18 December 2015 / Published online: 16 February 2017

Abstract Rating scales are used to elicit data about qualitative entities (e.g., research collaboration). This study presents an innovative method for reducing the number of rating scale items without the predictability loss. The "area under the receiver operator curve method" (AUC ROC) is used. The presented method has reduced the number of rating scale items (variables) to 28.57% (from 21 to 6) making over 70% of collected data unnecessary. Results have been verified by two methods of analysis: Graded Response Model (GRM) and Confirmatory Factor Analysis (CFA). GRM revealed that the new method differentiates observations of high and middle scores. CFA proved that the reliability of the rating scale has not deteriorated by the scale item reduction. Both statistical analysis evidenced usefulness of the AUC ROC reduction method.

✉ D. Strzałka
strzalka@prz.edu.pl

W. W. Koczkodaj
wkoczkodaj@cs.laurentian.ca

J. Montero-Marin
jmonteromarin@hotmail.com

[1] Computer Science, Laurentian University, 935 Ramsey Lake Rd., Sudbury, ON P3E 2C6, Canada

[2] Sudbury Therapy, Sudbury, ON, Canada

[3] UKSW University, Dewajtis 5, 01-815 Warsaw, Poland

[4] Faculty of Health Sciences and Sports, University of Zaragoza, Saragossa, Spain

[5] Centre for Global Mental Health, London School of Hygiene and Tropical Medicine, London, UK

[6] Miguel Servet Hospital, University of Zaragoza, Saragossa, Spain

[7] Jagiellonian University, Gołębia 24, 31-007 Kraków, Poland

[8] Faculty of Electrical and Computer Engineering, Rzeszów University of Technology, Al. Powstańców Warszawy 12, 35-959 Rzeszow, Poland

Keywords Rating scale · Prediction · Receiver operator characteristic · Reduction

Mathematics Subject Classification 94A50 · 62C25 · 62C99 · 62P10

Introduction

Rating scales (also called assessment scale) are used to elicit data about quantitative entities (e.g., research collaboration as in Bornmann et al. (2009)). Often, predictability of rating scales (also called "assessment scales") could be improved. Rating scales often use values: "1 to 10" and some rating scales may have over 100 items (questions) to rate. Other popular terms for rating scales are: *survey* and *questionnaire* although a questionnaire is a method of data collection while survey may not necessarily be conducted by questionnaires. Some surveys may be conducted by interviews or by analyzing web pages. Rating itself is very popular on the Internet for "Customer Reviews" where often uses five stars (e.g., by Amazon.com) instead of ordinal numbers. One may regard such rating as a one item rating scale. Surveys are used in Cinzia and Wolfgang (2016) on Fig. 1 (with the caption: "Sketch of data integration in use for different purposes with interference points for standardisation") as one of the main sources of data.

A survey, based on the questionnaire, answered by 1704 researchers from 86 different countries, was conducted by the Scientometrics study (Buela-Casal and Zych 2012) on the impact factor, which is regarded as a controversial metric. Rating scales were also used in Prpic (2007) and Koczkodaj et al. (2014). In Kakiashvili et al. (2012) and Gan et al. (2013), a different type of the rating scale improvement was used (based on pairwise comparisons). The evidence of improving accuracy by pairwise comparisons is in Koczkodaj (1996) and Koczkodaj (1998).

According to Moigne and Ragouet (2012):

> ... the differentiation of sciences can be explained in a large part by the diffusion of generic instruments created by research-technologists moving in interstitial arenas

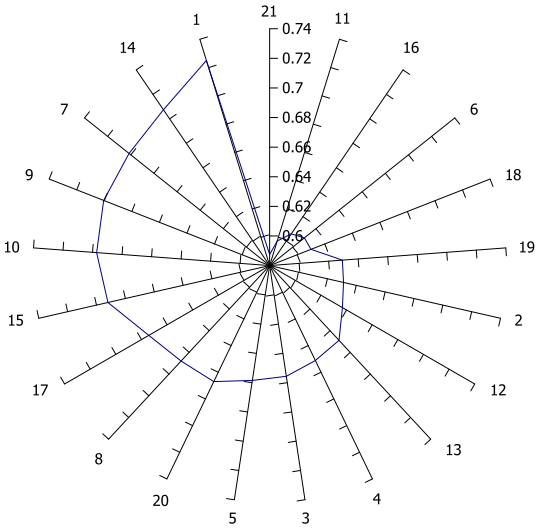

Fig. 1 AUC for the running total of all variables

between higher education, industry, statistics institutes or the military. We have applied this analysis to research on depression by making the hypothesis that psychiatric rating scales could have played a similar role in the development of this scientific field.

The absence of a well-established unit (e.g., one kilogram or meter) for measuring the science compels us to use rating scales. They have great application to scientometrics for measuring and analyzing performance based on subjective assessments. Even granting academic degrees is based on rating scales (in this case, several exams which are often given to students by questionnaires). Evidently, we regard this rating scale as accurate otherwise our academic degrees may not have much value.

The importance of subjectivity processing was driven by the idea of *bounded rationality*, proposed by Herbert A. Simon (the Nobel Prize winner), as an alternative basis for the mathematical modelling of decision making.

The data model

Data collected by a rating scale with fixed number of items (questions) are stored in a table with one decision (in our case, binary) variable. The parametrized classifier is usually created by total score of all items. Outcome of such rating scales is usually compared to external validation provided by assessing professionals (e.g., grant application committees).

Our approach not only reduces the number of items but also sequences them according to the contribution to predictability. It is based on the Receiver Operator Characteristic (ROC) which gives individual scores for all examined items Table 1.

Predictability measures

The term "receiver operating characteristic" (ROC), or "ROC curve" was coined for a graphical plot illustrating the performance of radar operators (hence "operating"). A binary classifier represented absence or presence of an enemy aircraft and was used to plot the fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate). Positive instances (P) and negative instances (N) for some condition are computed and stored as four outcomes a 2 contingency table or confusion matrix, as follows:

In assessment and evaluation research, the ROC curve is a representation of a "separator" (or decision) variable. The decision variable is usually: "has a property" or "does not have a property" or has some condition to meet (pass/fail). The frequencies of positive and negative cases of the diagnostic test vary for the "cut-off" value for the positivity. By changing the "cut-off" value from 0 (all negatives) to a maximum value (all positives), we obtain the ROC by plotting TPR (true positive rate also called sensitivity) versus FPR (false positive also called specificity) across varying cut-offs, which generate a curve in the unit square called an ROC curve.

According to Fawcett (2006), the area under the curve (the AUC or AUROC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming the 'positive' rank higher than 'negative').

Table 1 The confusion matrix

True positives	False positives
False negative	True negative

AUC is closely related to the *Mann-Whitney U test* which tests whether positives are ranked higher than negatives. It is also equivalent to the Wilcoxon test of ranks. The AUC is related to the Gini coefficient given by the formula

$$G_1 = 2 * AUC - 1, \tag{1}$$

where

$$G_1 = 1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1})$$

In this way, it is possible to compute the AUC using an average of a number of trapezoidal approximations. Practically, all advanced statistics can be questioned and they often gain recognition after their intensive use. The number of publications with ROC listed by PubMed.com has exploded in the last decade and reached 3588 in 2013. An excellent tutorial-type introduction to ROC is in Fawcett (2006). It was introduced during the World War II for evaluation of performance the radar operators. Its first use in health-related sciences, according to Medline search, is traced to Carterette and Jones (1967).

Validation of the predictability improvement

Supervised learning is the process of inferring a decision (of classification) from labeled training data. However, the supervised learning may also employ other techniques, including statistical methods that summarize and explain key features of the data. For the unsupervised learning, clustering is the most popular method for analyzing data. The *k*-means clustering optimizes well for the given number of classes. In our case, we have two classes: 0 for "negative" and 1 for "positive" outcome of diagnosis for depression.

The area under the receiver operating characteristic curve (AUC) reflects the relationship between sensitivity and specificity for a given scale. An ideal scale has an AUC score equal to 1 but it is not realistic in clinical practice. Cutoff values for positive and negative tests can influence specificity and sensitivity, but they do not affect AUC. The AUC is widely recognized as the performance measure of a diagnostic test's discriminatory power (see Lasko et al. 2005; Zou et al. 2007). In our case, the input data have AUC of 81.17%.

The following System R code was used to compute the AUC for all 21 individual items:

```
library(caTools)
# read data from a csv file
mydata = read.csv("C: \\BDI571.csv")
y = mydata[,1]
result¡-matrix(nrow=22,ncol=2);
ind=2;

for (i in 2:22)
{
      result[ind,]=colAUC(cbind(mydata[,1],
           mydata[,i]),y, plotROC=FALSE, alg="ROC")
      ind = ind+1
}
```

System R code

When AUC values are computed for all individual variables, we arrange them in an ascending order. These variables are present in Table 3 in bold. Values in the row below running total up to the current variable. Evidently, the first value 0.725 is the same as in Table 2 since the running total is the single variable 1. However, the third value in the second row (0.795) is not for variable 7 but the total of variables 1, 14, and 7. In particular, the last value (0.812) in Table 3 is for the total of all variables. Frankly, these numbers are very close to each other but their line plot 1 demonstrates its usefulness. The curve peek is for variable #6 which is 15. There is a slight decline until variable 16.

Relating the results to graded response model

Let us examine how our results can be related to the Graded Response Model (GRM). GRM is equivalent of Item Response Theory, well addressed by a Wikipedia article, but used for ordinary, not binary, data. GRM is usually conducted to establish the usefulness of test items (Ayala et al. 1992).

GRM is used in psychometric scales to determine the level of three characteristics of each item, namely: (a) item's difficulty, (b) item's discriminant power, and (c) item's guessing factor.

Item's difficulty describes how difficult or easy it is for individuals to answer on the item. High positive value means that the item is very difficult, high negative value means that the item is very easy.

Item's discriminant power describes ability for a specific item to distinguish among upper and lower ability individuals' on a test.

Item's guessing factor describes probability that individual with low feature (low depression) achieved high scores in this item.

Table 2 AUC of individual variables in the original data

Var	AUC	Var	AUC	Var	AUC
21	0.587468	12	0.636791	17	0.674283
11	0.597342	13	0.648917	15	0.692064
16	0.605937	4	0.651187	10	0.697225
6	0.610004	3	0.655666	9	0.700461
18	0.610028	5	0.658478	7	0.701489
19	0.629285	20	0.666999	14	0.707401
2	0.631205	8	0.667983	1	0.725009

Table 3 AUC of running variable totals

1	**14**	**7**	**9**	**10**	**15**	**17**
0.725	0.777	0.795	0.810	0.813	0.822	0.821
8	**20**	**5**	**3**	**4**	**13**	**12**
0.819	0.820	0.821	0.821	0.821	0.821	0.820
2	**19**	**18**	**6**	**16**	**11**	**21**
0.819	0.818	0.816	0.814	0.812	0.811	0.812

The aim of our analysis was to establish whether or not the GRM indicates the same items as the proposed method based on AUC. Two GRM models were build for the given rating scales:

Constrained (that assumes equal discrimination parameters across items),

Unconstrained (that assumes unequal discrimination parameters across items).

System R *ltm* package (Rizopoulos 2006) was used in our analysis. Fig. 2 illustrates system R code for GRM models.

In order to check whether or not the unconstrained GRM provides a better fit than the constrained GRM, a likelihood ratio test was used. It revealed that unconstrained GRM is preferable (fit2 in Table 4). The results of the Likelihood Ratio are presented in Table 4.

Table 5 shows the unconstrained GRM model results with the item discrimination power. It provides information on discrimination power of each item.

Items selected by AUC ROC are shown in Table 5 as bold. Evidently, they have the large discrimination power (seen in the last column). All selected items discriminate between responses above the mean value (so on their basis we can discriminate between respondents with severe and moderate level of depression). Discrimination power is a characteristic of items in the scale. It is a measurement method which aim is to assess how respondents differ in their answers on rating scale items. The larger is the discrimination power of the item, the better, more useful is item in the scale (Anastasi and Urbina 1999). Items computed by the proposed (AUC ROC) method have a good discrimination as it can be seen in the Table 4 (for example, number 1.799 means that item *V*1 has a good discrimination power).

All items of the given rating scale give 56.21% of total information for the latent trait and the latent variable (adolescent depression in school in our case). Test Information Curve (see Table 6) shows that six items provides 19.62% of the total information for latent trait. The higher is items' discrimination, the more information or precision the scale provides.

GRM model computes different items than our proposed method. AUC ROC is based on the count of true and false positive rate while GRM model is based on the maximum likelihood estimate. The proposed method has a bigger diagnostic power. Diagnostic power is the ability of the test to detect all subjects, which have been measured by the test characteristics (in our case, for depression). A test with the maximum diagnostic power would detect all subjects (suffering from depression). Unfortunately, the most selections of rating scale items do not compare solutions with the diagnostic criterion. That is why the

Fig. 2 System R code for GRM models

```
fit <-grm (BDI571,constrained=TRUE)
grm (BDI571,constrained=TRUE)

fit2 <-grm (BDI571)
grm (BDI571)

anova (fit1,fit2)
```

Table 4 Likelihood ratio for the full GRM model

	AIC	BIC	log.Lik	LRT	*df*	*p* value
Fit1	25494.12	25772.35	−12683.06			
Fit2	25367.63	25732.81	−12599.81	166.49	20	$p < 0.001$

Table 5 Unconstrained GRM model results for the full rating scale and the item discrimination power

	Extrmt1	Extrmt2	Extrmt3	Dscrmn
V1	**0.178**	**1.099**	**2.542**	**1.799**
V2	0.214	1.536	2.485	1.315
V3	–0.094	1.306	3.077	1.528
V4	–0.447	1.673	3.511	1.268
V5	–0.709	1.903	2.717	1.440
V6	–0.073	1.679	2.194	0.860
V7	**–0.410**	**0.970**	**2.170**	**1.459**
V8	–0.641	0.876	2.157	1.461
V9	**0.092**	**1.895**	**2.471**	**1.405**
V10	**–0.183**	**0.834**	**1.665**	**1.023**
V11	–0.881	2.187	3.280	0.767
V12	–0.242	1.282	2.221	1.271
V13	–0.351	1.631	2.660	1.054
V14	**–0.038**	**0.918**	**2.353**	**1.951**
V15	**–0.627**	**1.046**	**2.248**	**1.593**
V16	–2.364	0.634	2.388	0.764
V17	–0.482	1.287	2.366	1.296
V18	–1.685	0.847	2.024	0.902
V19	–1.623	0.366	2.648	1.078
V20	–0.575	1.227	2.066	1.643
V21	1.271	2.240	3.531	0.870

Table 6 Test information curve

Total information = 56.21
Information in (−4, 4) = 52 (92.51%)
Based on all the items
Total information = 19.62
Information in (−4, 4) = 18.97 (96.65%)
Based on items 1, 7, 9, 10, 14, 15

proposed method is so useful for the selection of items in different measurement tools (examination, tests, socio-metrical scales, psychometrical scales, and many others).

We used GRM model here to show that even such powerful method like GRM (used in psychometrics to indicate which items can discriminate subjects), does not provide an answer to a question about diagnostic accuracy of items. According to GRM items, V2 and V3 (Table 5) have a considerable discriminant power, but the proposed method shows which items better discriminate between subjects on the basis of diagnostic criteria.

Reduced scale psychometric properties

Confirmatory Factor Analysis (CFA) (Hair et al. 2006; Bartholomew et al. 2008) was used to verify the structure of our results. CFA is a factor analysis which purpose is to verify the structural validity whether items belong to scales and what are their factor loading. Factor

```
> HS.model<-depression='V1+V7+V9+V10+V14+V15'
> fit<-cfa(HS.model,data=BDI571,ordered=c("V1","V7","V9","V10","V14","V15"))
> summary(fit,fit.measures=TRUE,standardize=TRUE)

> HS.model<-depression='V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11+V12+V13+V14+V15+
V16+V17+V18+V19+V20+V21'
> fit<-cfa(HS.model,data=BDI571,ordered=c("V1","V2","V3","V4","V5","V6",
"V7","V8","V9","V10","V11","V12","V13","V14","V15","V16",
"V17","V18","V19","V20","V21"))
> summary(fit,fit.measure=TRUE,standardize=TRUE)
```

Fig. 3 System R code for CFA

loading measures the relations between observed variable (item) and latent feature (scale). The higher the factor loading, the stronger the relation, and the item has greater importance in the scale. More specifically, CFA was used to determine whether:

– items indicated by AUC form a coherent scale that exhibits good reliability,
– the reliability of the rating scale has not deteriorated by the scale item reduction.

Two CFA models were built. The first CFA model has all items and the second CFA model has a reduced number of items. Since items of the scale have categorical format, the robust estimator WLSMV (weighted least squares means and variance, see Beauducel and Herzberg (2006)) was used as it is designed for categorical scales. The robust estimator resists the lack of normal distributions. The analysis was conducted in "lavaan" package of R program (Fig. 3).

The model for the full rating scale is presented by Fig. 4. Table 7 presents parameter estimates of the full rating scale. Loads of those items, which have been identified by the presented method as having the greatest predictive power, is in bold in Table 7. A model with a reduced number of items is in Fig. 5. Table 8 presents parameter estimates for the reduced scale model.

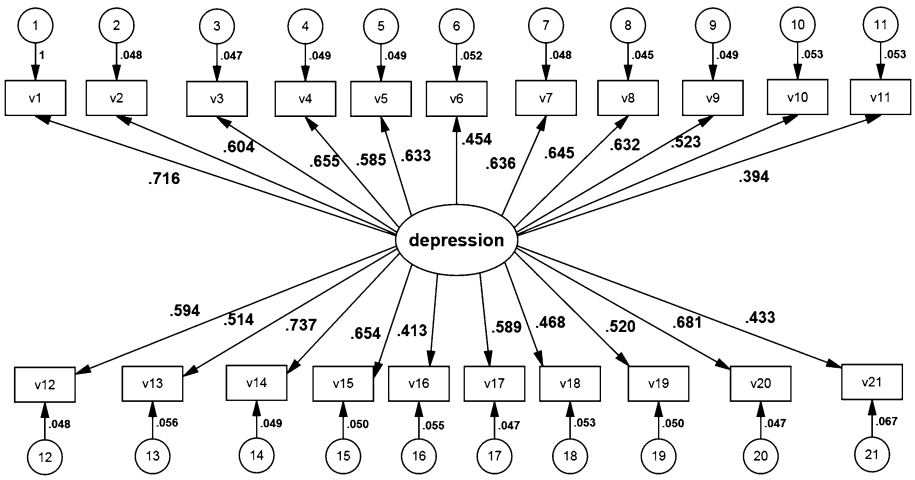

Fig. 4 CFA model for the rating scale with all items presented in AMOS graphics

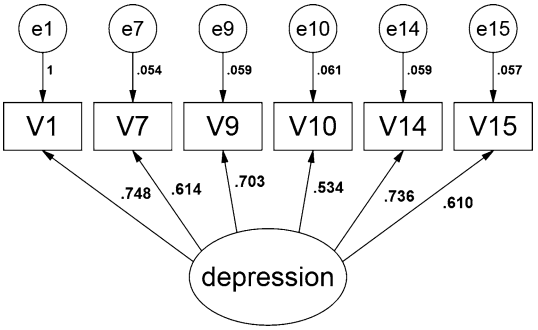

Fig. 5 CFA Model with a reduced number of items presented in AMOS graphics

Table 7 Parameter estimates of the full rating scale

Parameters	Standardized	Non-standardized	Standardized error
λ_{V1}	**0.716**	1.000	
λ_{V2}	0.604	0.844	0.048
λ_{V3}	0.655	0.914	0.047
λ_{V4}	0.585	0.818	0.049
λ_{V5}	0.633	0.884	0.049
λ_{V6}	0.454	0.634	0.052
λ_{V7}	**0.636**	0.889	0.048
λ_{V8}	0.645	0.901	0.045
λ_{V9}	**0.632**	0.883	0.049
λ_{V10}	**0.523**	0.731	0.053
λ_{V11}	0.394	0.550	0.053
λ_{V12}	0.594	0.830	0.048
λ_{V13}	0.514	0.718	0.056
λ_{V14}	**0.737**	1.029	0.049
λ_{V15}	**0.654**	0.913	0.050
λ_{V16}	0.413	0.578	0.055
λ_{V17}	0.589	0.822	0.047
λ_{V18}	0.468	0.653	0.053
λ_{V19}	0.520	0.726	0.050
λ_{V20}	0.681	0.952	0.047
λ_{V21}	0.433	0.605	0.067

Table 8 Parameter estimates for the reduced scale

Parameters	Standardized	Non-standardized	Standardized error
λ_{V1}	0.748	1.000	
λ_{V7}	0.614	0.821	0.054
λ_{V9}	0.703	0.940	0.059
λ_{V10}	0.534	0.714	0.061
λ_{V14}	0.736	0.984	0.059
λ_{V15}	0.816	0.816	0.057

For the purpose of checking whether the models have a good fit, we used two fit indices: CFI (cross validation index) and RMSEA (root mean square error of approximation). According to Bartholomew et al. (2008) and Saris et al. (2009), both CFA models have a good fit to the data as illustrated by Table 9. Values of CFI statistics for both models exceeded the required level of 0.9. For both models, the values of RMSEA statistics (lower than 0.08) indicates the good fitness of the proposed new scale structure for the given data.

For both CFA models, *construct reliability* (CR) and *variance extracted* (VE) were computed. *CR* was computed by the formula (given in Hair et al. (2006)):

$$CR = \frac{\left(\sum_{i=1}^{n} \lambda_i\right)^2}{\left(\sum_{i=1}^{n} \lambda_i\right)^2 + \left(\sum_{c=1}^{n} \delta_i\right)} \tag{2}$$

where i is a total number of items, λ is a factor loading, δ is an error variance, which is the amount of variability unexplained by the items in scale.

The formula for computing *variance extracted* (VE) is based on Hair et al. (2006):

$$VE = \frac{\sum_{i=1}^{n} \lambda_i^2}{n} \tag{3}$$

where i is the number of items, λ is a factor loading, n is a number of rating scale items.

The results revealed that the reliability of the reduced model CR = .822 and is lower than the reliability of the full model of 0.1 (CR =.929). Therefore, it can be concluded that the reliability of the scale is above the acceptability level. Removing 15 items has not impaired its reliability as Table 10 demonstrates it.

For the reduced model, VE = .438 while for the given model, VE = .394. Evidently, the new model has VE closer to criterion of .500. The reduced rating scale model has a better VE than the full rating scale model. It means that the reduced rating scale model explains the diversity of the results better than the full rating scale model (see Table 10).

On the basis of factor loadings (λ), we are unable to determine which items have the most predictive power. Items V3 or V20 have one of the top factor loadings in the full rating scale, but they do not still have the most predictive power. Therefore, it is impossible to indicate the ordinal number of the rating scale item according to the factor analyses, but it is possible by the proposed method and GRM. However, GRM cannot compare its solution with a diagnostic criterion while the proposed method can.

Table 9 Results of fit statistics for two rating scale models

Statistics for the full and reduced rating scale models	
Chi_2 = 437.899	Chi_2 = 30.883
df = 189	*df* = 9
CFI = .950	CFI = .983
RMSEA = .048	RMSEA = .065

Table 10 Results of CR and VE of two models

Rating scale	CR	VE
Full rating scale	0.929	0.394
Reduced scale	0.822	0.483

Discussion

The Beck Depression Inventory (BDI) was selected for our study since it is one of the best known and most widely used self-rating scales to assess the presence and severity of depressive symptoms (Beck et al. 1996; American Psychiatric Association 2000; World Health Organization 1994). Our data were collected in high schools (Araya et al. 2011, 2013). However, it needs to be stressed that our method is applicable to practically all rating scales.

In summary, both models fit the data well. Both of them have a good reliability and a relatively good variance. Reducing the number of items did not burden psychometric properties, but simplified the whole structure (as indicated by the smaller number of degrees of freedom). According to the Occam's Razor law, the simpler models, the better. Although it was not the main objective of this study, it is worth to notice that the six rating scale items have a better predictive power in our study than other 21 items. We have also demonstrated that our results have the domain (semantic) meaning.

Conclusions

The presented method has reduced the number of the rating scale items (variables) to 28.57% (from 21 to 6) making over 70% of collected data unnecessary. It is not only an essential budgetary saving, as data collection is usually expensive, but it often contributes to the data collection error reduction. The more data are collected, the more errors are expected to occur. When we use the proposed AUC ROC reduction method, the predictability has increased by approximately 0.5%. It may seem insignificant but for a large population, it is not so. In fact, http://www.who.int/mental_health/publications/action_plan/en/ states that: "Taken together, mental, neurological and substance use disorders exact a high toll, accounting for 13% of the total global burden."

The proposed use of AUC for reducing the number of rating scale items is innovative and applicable to practically all rating scales. System R code is posted on the Internet for the general use. A package for System R is under development. Certainly, more validation cases would be helpful and the assistance will be provided to anyone who wishes to try this method using his/her data.

Supporting information

The source code will be deposited at SourceForge.net hosting provider (see http://www.sourceforge.net/). According to http://www.sourceforge.net/, SourceForge "creates powerful software in over 400,000 open source projects ans hosts over 3.7 million registered users". It connects well over 40 million customers with more than 4,800,000 downloads a day.

Acknowledgements The first author was supported (in part) by the Euro Research grant "Human Capital". Authors would like to thank Prof. E. Aranowska, specialized in psychometry, for reading the first draft. Authors would like to thank Amanda Dion-Groleau (Laurentian University, Psychology) and Grant O. Duncan, Team Lead, Business Intelligence and Software Integration, Health Sciences North, Sudbury, Ontario for their help with proofreading this text. The presented method is implemented as "R package" (submitted to the Comprehensive R Archive Network, CRAN), named RatingScaleReduction.

References

Anastasi, A., & Urbina, S. (1999). *Testy psychologiczne*. Warszawa: Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.

Araya, R., Montero-Marin, J., Barroilhet, S., Fritsch, R., & Montgomery, A. (2013). Detecting depression among adolescents in Santiago, Chile: Sex differences. *BMC Psychiatry*, *13*, 269.

Araya, R., Montgomery, A. A., Fritsch, R., Gunnell, D., Stallard, P., Noble, S., et al. (2011). School-based intervention to improve the mental health of low-income, secondary school students in Santiago, Chile (YPSA): study protocol for a randomized controlled trial. *Trials*, *12*, 49.

Ayala, R. J. D., Dodd, B. G., & Koch, W. R. (1992). A comparison of the partial credit and graded response model in computerized adaptive testing. *Applied Measurement in Education*, *5*(1), 17–34.

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2008). *Analysis of multivariate social science data*. Boca Raton, FL: Chapman & Hall/CRC Press.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA structural equation. *Structural Equation Modeling: Modeling A Multidisciplinary Journal*, *13*(2), 186–203.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II - beck depression inventory manual* (Vol. Second). San Antonio: The Psychological Corporation.

Bornmann, L., Mutz, R., & Daniel, H. D. (2009). The influence of the applicants' gender on the modelling of a peer review process by using latent Markov models. *Scientometrics*, *81*(2), 407–411.

Buela-Casal, G., & Zych, I. (2012). What do the scientists think about the impact factor? *Scientometrics*, *92*(2), 281–292.

Carterette, E. C., & Jones, M. H. (1967). Visual and auditory information processing in children and adults. *Science*, *156*(3777), 986–988.

Cinzia, D., & Wolfgang, G. (2016). Grand challenges in data integrationstate of the art and future perspectives: An introduction. *Scientometrics*, *108*, 391–400.

Diagnostic and Statistical Manual of Mental Disorders. (2000). text revision (DSM-IV-TR), 4th Revision. American Psychiatric Association.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Gan, Y., Kakiashvili, T., Koczkodaj, W. W., & Li, F. (2013). A note on relevance of diagnostic classification and rating scales used in psychiatry. *Computer Methods and Programs Biomedicine*, *112*(1), 16–21.

Hair, J. J., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*. New Jersey: Upper Saddle River.

ICD-10. (1994). *International Classification of Diseases*. Geneva: World Health Organization.

Kakiashvili, T., Koczkodaj, W. W., & Woodbury-Smith, M. (2012). Improving the medical scale predictability by the pairwise comparisons method: Evidence from a clinical data study. *Computer Methods and Programs Biomedicine*, *105*(3), 210–216.

Koczkodaj, W. W. (1996). Statistically accurate evidence of improved error rate by pairwise comparisons. *Perceptual and Motor Skills*, *82*, 43–48.

Koczkodaj, W. W. (1998). Testing the accuracy enhancement of pairwise comparisons by a Monte Carlo experiment. *Journal of Statistical Planning and Inference*, *69*(1), 21–31.

Koczkodaj, W. W., Kulakowski, K., & Ligeza, A. (2014). On the quality evaluation of scientific entities in Poland supported by consistency-driven pairwise comparisons method. *Scientometrics*, *99*(3), 911–926.

Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, *38*(5), 404–415.

Le Moigne, P., & Ragouet, P. (2012). Science as instrumentation. *The Case for Psychiatric Rating Scales Scientometrics*, *93*(2), 329–349.

Prpic, K. (2007). Changes of scientific knowledge production and research productivity in a transitional society. *Scientometrics*, *72*(3), 487–511.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling A Multidisciplinary Journal*, *16*, 561.

sourceforge.net/. Accessed 14 Dec 2015.

World Health Organization. (2013). Mental Health Action Plan 2013 - 2020, WHO Library Cataloguing-in-Publication Data http://www.who.int/mental_health/publications/action_plan/en/ Accessed 1 May 2015.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *115*(5), 654–657.