Survey Research Methods (2018)
Vol. 12, No. 2, pp. 147-160
doi:10.18148/srm/2018.v12i2.7224

ISSN 1864-3361
http://www.surveymethods.org

How to Avoid the Sins of Questionnaires Abridgement?

Paweł Kleka
Adam Mickiewicz University in Poznan
Department of Social Science
Institute of Psychology

Emilia Soroko
Adam Mickiewicz University in Poznan
Department of Social Science
Institute of Psychology

The creation of abridged versions of research tools is a common, justifiable process, but unfortunately it is often carried out without due methodological care and regard for the consequences. Smith and collaborators (2000) have already written about the mistakes that can be made, but their article has not had much practical impact. There are two main mistakes commonly made by researchers: assuming the transferability of validity and reliability between the full and shortened versions and using less stringent criteria to assess the validity and reliability of short forms. These two problems manifest as nine sins committed during the construction of short forms. Here we present procedures designed to allow researchers to avoid these mistakes and create abridged versions of research tools that are as reliable as possible and to assess the costs of the various methods of abridging questionnaires. To this end we determine the expected length of the tool and weight the benefits of reduced questionnaire completion time against the loss of reliability. We also estimate the shared variance of the full and short versions and classification accuracy of the new, short version. We compared quality of short form obtained from three most common statistical techniques for abridging questionnaires. We analysed data from a sample of 519 persons; 309 (59.5%) completed the paper version of the Self-Narrative Inclination Questionnaire (IAN-R), and 210 (40.5%) participated in online tests. Based on the analyses of difficulty and discriminatory power we found that abridgements built on item response theory analyses had a slight advantage over abridgements based on factor loadings and Cronbach's α.

Keywords: test abridgement guidelines; shortening of test; abbridged version; reliability

1 Introduction

Use of short forms (SF) of measurement tools in psychology can be practically and theoretically justifiable. The main reasons for using shortened version are to reduce test administration time, to ensure that the length and effort required are suited to the capabilities of respondents and to reduce the cost of test administration. For example, one may need to use SFs of tests when a research plan involves administering a battery of tests, because administering the full forms (FFs) would impose an excessive burden on subjects. Another potential benefit of using SFs may is that administration time is better matched to subjects' cognitive and emotional capabilities, thus reducing the influence of these factors on the measurement of the variables of interest.

Psychologists who abridge questionnaires may be tempted to create SFs intuitively without regard for the integrity and reliability of the result or, contrarily, to base a SF exclusively on the results of mathematical calculations. New SFs of tools are not always subjected to rigorous psychometric analysis.

A review of the literature on abridgement uncovered several publications from various sub-disciplines of psychology that have reported on SFs and discussed researchers' decisions (e.g. questionnaires: Coroiu et al., 2015; Las Hayas, Quintana, Padierna, Bilbao, and Muñoz, 2010; Manos, Kanter, and Luo, 2011; Schaufeli, Bakker, and Salanova, 2006; tests: Arthur and Day, 1994; Van der Elst et al., 2013). Some authors have raised the issue of the quality of SFs (for example: Bersoff, 1971; Fox, McManus, and Winder, 2001; Hamel and Schmittmann, 2006; van Dierendonck, Díaz, Rodríguez-Carvajal, Blanco, and Jiménez, 2008). There has also been discussion on abridgement of diagnostic tests, mainly in the field of intellectual deficits assessment, which had led to some reflection on the merits of various abridgement techniques in relation to the structure of the original (Bersoff, 1971; Kaufman, 1972; Satz & Mogel, 1962; Silverstein, 1990; Warrington, James, & Maciejewski, 1986). A distinction has also been drawn between abridgement for diagnostic purposes and abridgement for research and screening purposes (Bors & Stokes, 1998; Clara & Huynh, 2003;

Contact information: Paweł Kleka, ul. Szamarzewskiego 89B; 60-568 Poznań (E-Mail: pawel.kleka@amu.edu.pl).
Replication materials are available in the supplementary files, and also on https://osf.io/5k9cx/.

Crawford, Allan, & Jack, 1992; Hamel & Schmittmann, 2006). There have, however, been few attempt to produce general, theoretically based, practical guidance on how to carry out abridgement (see also Smith, McCarthy, and Anderson, 2000; for an example in the higher education field see Lee, Bygrave, Mahar, and Garg, 2014). An empirical comparison of shortening methods has not been supported by many literature sources yet (Kleka, 2013; Prieto, Alonso, & Lamarca, 2003; Ziegler, Kemper, & Kruyen, 2014).

Although abridgement of psychological questionnaires and tests is widely practiced there has so far been little published discussion about methodologically sound approaches to the process (see also Prieto et al., 2003). There are two main issues to consider. The first is the development of a SF: can one establish general guidelines for the process and recommend suitable statistical techniques for selecting items? The second is how SFs should be assessed. This article addresses both issues, taking into account the 'sins' described by Smith et al. (2000). Our most important innovation is the formulation of general guidelines for the abridgement procedure and we provide an illustration of their use. We would like to highlight the practical aspect of shortening, what is an extension of inspiring but theoretical work by Smith et al. (2000).

1.1 Abridgement Techniques based on Classical Test Theory and Item Response Theory

There are two main approaches to abridgement of psychological tools: 1) abridgement based on the internal homogeneity of the tool and 2) abridgement on the basis of item response theory (IRT). The abridgement of tools based on their internal homogeneity is rooted in classical test theory (CTT) and relies on covariance analysis. The main statistical techniques used in this approach are factor analysis, correlations between scores on the short and full versions, item-total correlations, Cronbach's reliability coefficient and stepwise regression coefficients (Coste, Guillemin, Pouchot, & Fermanian, 1997). All these procedures are based on internal homogeneity of a short version items (Nunnally & Bernstein, 1994). In abridging a tool solely based on internal homogeneity analysis a researcher the risk that scores on the SF will not have the same meaning as scores on the FF, i.e. that the SF will not measure the same underlying construct as the FF. This risk arises because items that are relatively weakly correlated with the overall score or with core items capture a peripheral component of the construct that the FF measures, and in excluding these items from the SF one is failing to measure the full construct. It can be reduced by using regression analysis, which provides insight into the structure of an instrument and allows researchers to select items based on criterion validity as well as internal homogeneity (Schipolowski, Schroeders, & Wilhelm, 2014). The abovementioned analytical techniques, based on CTT, do not assume any order of items or their possible hierarchy. SFs prepared on the basis of these techniques are characterized by greater internal homogeneity than the original (see also Prieto et al., 2003). These techniques result in preferential selection of items with a lot of common variance, which has an impact on the content validity of the SF and reduces the range of contexts in which it can be used (e.g. for screening but not diagnosis).

IRT-based methods address these limitations. They enable the items of a given tool to be ordered in terms of difficulty or discriminatory power and also indicate how well particular items describe respondents with various levels of the target variable. It is possible to determine a level of difficulty (theta) for each item of the FF, which enables selection of items covering the full spectrum of difficulty (advisable in construction of a SF to be used for research purposes) or from a selected critical region (advisable for screening purposes). When an IRT-based approach is used the researcher's selection decisions are based on knowledge about the internal structure of the FF, its dimensionality and the variables which may affect the results. Another simplification is offered through additivity, which facilitates the selection of items for SF, without a necessity to calculate item properties again after changing the content of a given version.

As well as the considering the technique used to select items for a SF, special attention should be given to a technique for assessing its predictive validity based on cross-validation. This involves assessing the extent to which the SF prepared on the basis of one sample (the training sample) predicts scores on the FF in another sample (the testing sample). When the results from one sample are generalized to other samples, such validation allows determining a degree of shrinkage of the predictive power of a predictive tool. It should be noted that this validation method allows one to state empirically "how much of power has been lost", but it does not indicate the quality of a SF. Kerlinger and Pedhazur (1973) have claimed that this is a very conservative method (*"the most rigorous approach to the validation of results from regression analysis in a predictive framework"*; Kerlinger and Pedhazur, 1973, p. 284). Cross-validation has made little impact on the literature on SFs of measurement tools and where it is used it correlations between scores on the SF and FF have been low and the standard errors high (Woo-Sam & Zimmerman, 1973). More detailed descriptions of the technique can be found on many websites[1].

Our choice of statistical techniques was dictated largely by practical considerations. The analyses we recommend can be performed in popular statistical packages such as SPSS, Statistica, Stata, etc., which most social science researchers are capable of using, but it is worth mentioning that there are other methods of constructing SFs using more complicated

[1]For example: https://en.wikipedia.org/wiki/Cross-validation_(statistics)

– in computing terms – techniques. One such technique is to use a genetic algorithm (GA) that 'evolves' good solutions based on evolutionary principles. This method is based on reducing the redundancy within a tool, i.e. reducing it to the subset of items that best capture the trait(s) of interest. Yarkoni (2010) demonstrated that the technique could be profitably applied, especially in contexts where the dimensionality of a construct is too high to afford an analytical solution. In one study Yarkoni generated a 181-item instrument that accurately recaptured the variance in 8 different scales consisting of a total of 2,019 items, representing a 91% reduction in instrument length. Another more computationally sophisticated abridgement technique uses the ant colony optimization (ACO) algorithm, which is based on the example of ants which mark the shortest way from their colony nest to food source with pheromones (Leite, Huang, & Marcoulides, 2008).

1.2 Major Psychometric Errors of Abbriding Psychological tools

It is rightly emphasized that the psychometric properties of a SF should not be substantially worse than those of the original (see also Choynowski, 1971). Smith et al. (2000) provided an excellent guide to the principles of abridgement in which they devote a lot of attention to describing the 'sins' psychologists commit when abridging psychometric tools. According to Smith et al. these sins derive from two unjustified assumptions.

The first mistaken assumption is that the validity and reliability of the original guarantees the validity and reliability of a shorter tool derived from it (Smith et al., 2000). The SF is a new tool and it must be treated as such in psychometric terms. Its properties remain in certain relation to the properties of FF but this relation is subjected to varied influences, connected with the content of a tool item.

The second source of potential errors is widespread agreement regarding lower validity and reliability of shortened tool, only due to the fact that it is shorter (Smith et al., 2000). It is difficult to understand why inferior research tools should be accepted simply because they are SFs of another tool that is of acceptable quality. There are some SFs that would face short shrift if they were evaluated as independent tools rather than as the SF of a recognized tool. In such cases providing information about abridgement serves a protective function, allowing SFs of dubious quality, whose psychometric properties have not been properly assessed to be applied in research.

Based on their review of reports on abridgement of various tools Smith et al. compiled a list of errors that resulted from taking the two above-mentioned false assumptions: 1) Development of a SF that is not properly validated; 2) Failure to show that SF preserves the coverage of all factors measured by the original; 3) Failure to show that SF measures each factor reliably, 4) Failure to show that SF shares sufficient variance with FF, using independent administrations, 5) Failure to show that SF reproduces the factor structure of a multifactorial original; where SF has only overall factors, failure to show that it preserves the content domains represented by the subfactors; 7) Failure to demonstrate the validity of all the factors included in the SF in an independent sample; 8) Failure to demonstrate that classification based on SF is sufficiently accurate; 9) Failure to show that SF offers meaningful savings in time or resources for the loss of validity.

How can one avoid these mistakes? We think that creation of SFs should be most of all justified and rewarding. Preparing SF of research questionnaire can be assessed by determining target parameters of the tool before constructing SF (a priori). This would allow one to carry out a cost-benefit analysis. After proving worth of abridging, but before embarking on the process, researchers should reassure themselves that the tool they propose to abridge possesses adequate psychometric properties. It is also worth considering the intended application(s) of the SF as this may affect the choice of items. If a test is to be used for screening (in which case it will usually be administered to groups) the cut-off point is important and the procedure for choosing items will be different from the procedure used if a test is to be used for diagnosis (when it will usually be administered to individuals and it is important that the tool is capable of measuring as wide a range as possible of the variable of interest). The parameters used to assess the quality of the SF and make a decision about the extent of abridgement are as follows (cf. Smith et al., 2000)) 1) a priori SF reliability; 2) percentage of variance common to FF and SF; 3) validity as a priori correlation of SF with key criteria and (if possible) 4) a priori classification accuracy. Another important parameter is 5) the expected savings in time and/or resources as a result of using the SF.

2 Research Goals and Problems

The goals of this study were to put forward a procedure for abridging research questionnaires; show an empirical evaluation of abridgement techniques and a method of determining the psychometric properties of a SF. The need to create SF resulted from practical reasons, and this study is intended to provide a guide to abridging research questionnaires. We start by describing the research procedure and tool properties (contextualization), move on to calculation of an acceptable length for the SF and then describe the results of comparison of three statistical techniques for selecting items for a SF.

This brings us to an additional research question, do the various analytical techniques used for abridging questionnaires give similar results and are they equally useful to psychologists seeking to develop abridged versions of tools? We used the three most popular statistical techniques to construct short versions of the IAN-R. The first was based on IRT

and involved use of a polytomous Graded Response Model (GRM) (Samejima, 1969). The second was based on reliability analysis (RTT) of items which were part of the FF, using Cronbach's and Guttman's (Guttman, 1945). The third technique was based on the size of factor loadings determined using exploratory factor analysis (EFA). The section on abridgement procedure finishes with the calculation of the psychometric properties of the SF in our example, including classification accuracy.

We those not to use confirmatory factor analysis (CFA) because 1) it requires a larger sample than the other methods we used and makes more stringent assumptions about the sphericity of variables (Konarski, 2014); 2) Bartholomew (1987), Takane and De Leeuw (1987) independently demonstrated that 2PL models in Item Response Theory (IRT) and CFA models of discrete variables (the vast majority of responses to questionnaires are such variables) are formally equivalent. Moreover, as CTT and IRT are the most common psychometric approaches, we wanted compare their effectiveness. It proved relatively easy to identify the most appropriate IRT model, but this was far from the case with CTT, so we chose the most frequently used techniques: factor analysis (to determine factor loadings) and its rival, regression analysis (to determine correlation coefficients).

3 Example of Scale Abridgement: Creation of a Short Form of a Questionnaire of Self-Narrative Inclination (IAN-R)

3.1 Participants

Data were collected from persons (Table 1); 309 (59.5%) completed the paper version of the IAN-R and 210 (40.5%) participated in online tests. The format in which questionnaire was completed did not affect the results (scale M: $W = 30086.5$, $z = 0.22$, $p = 0.825$; scale D: $W = 30477.0$, $z = 1.33$, $p = 0.183$; scale K: $W = 32210.0$, $z = 0.74$, $p = 0.458$.

The groups completing the tests in each format (paper and pencil; Internet) differed with respect to educational level. The majority of the pencil-and-paper group (259 out of 332, 78%) had completed secondary education whereas the majority of the Internet group had a higher education qualification. Women made up the majority of both groups (overall F/M = 2.6458).

Considering just the participants educated to secondary school level, the paper-and-pencil was about 8 years younger than the Internet group; $F(1, 320) = 84.82$, $p < 0.001$, $\eta^2 = 0.21$. Considering just the participants with higher education the Internet group was about 10 years older $F(1, 173) = 41.31$, $p < 0.001$, $\eta^2 = 0.19$. Gender were equal in age except in the subset of the Internet group who had higher education, a *post hoc* Tukey's HSD test indicated that the men were about 5 years older than the women $d = 5.02$, $CI_{95} = [0.24, 9.80]$, $p = 0.035$.

After constructing the final short version of the IAN-R we analysed its reliability and validity in a new sample ($N = 177$). This sample was similar to the sample used to develop the SF in terms of age ($M = 29.6$ years, $SD = 11.29$), and education (median = secondary), but containing a higher proportion of women ($N_F = 139$, gender ratio F/M= 3.66).

Because multivariate outliers distort the results of correlation analyses, participants ($n = 16$) who showed a Mahalanobis distance with a χ^2 value significant at $p < 0.001$ (Tabachnick & Fidell, 2001) were excluded from further analysis so the final samples count 503 participants. Form of participation does not differentiate results ($t(478) = 1.09$, n.s.) or gender ($t(187) = 0.76$; n.s.), and so subsequent analyses were collapsed across gender and format. Finally, we examined the reliabilities of all scale scores. All scores hed satisfactory Cronbach's $\alpha > 0.70$ (Table 2).

3.2 Material

We demonstrated the abridgement of questionnaires using a test of self-narrative inclination (IAN-R; Soroko, 2013, 2014)[2]. The IAN-R is used to measure individual differences in aspects of self-narrative activity. It is a typical multidimensional (three moderately correlated scales) medium-length questionnaire (30 items) which measures narrative recounting (M)[3], distancing from experience (D)[4] and tendency to use cultural heritage to understand oneself (K)[5]. Responses are given using a five-point Likert scale ranging from 1 (agree) to 5 (disagree) and the questionnaire takes about 15 – 20 minutes to complete. The questionnaire can be used as an independent tool measuring self-narrative inclination, but also controlling the inclination in the subjects participating in qualitative interviews.

3.3 Procedure

Our first step towards creation of a SF of the IAN-R questionnaire was to divide the sample into a training group (80% of the sample) and a testing group (the remaining 20%). Training group were used for evaluation of particular analytical methods and three SFs containing the items shown in Table 2. Next, the testing group was used to analyze the quality of the three SFs.

[2]Full Polish and English version are avaiable in the supplementary files and ad at https://osf.io/5k9cx/

[3]Inclination to recount autobiographical events in a convention of a story in which the narrator is a protagonist.

[4]Inclination to take third -party perscpective on one's life and speculate out events involving oneself, in particular about the impact of past events on the present.

[5]Inclination to make use of cultural heritage (e.g., books, films, cultural patterns) when talking about one's own experiences.

Table 1
Age in groups according to sex, research type and education

Sex	Group	Education	*N*	*M*	SD	Md	Min	Max	Skew	Kurtosis	SE
F	I	basis	3	16.7	2.31	18.0	14	18	−0.385	−2.333	1.333
M	I	basis	2	15.5	2.12	15.5	14	17	0.000	−2.750	1.500
M	P	basis	1	23.0	-	23.0	23	23	-	-	-
F	I	vocational	1	48.0	-	48.0	48	48	-	-	-
M	I	vocational	3	40.0	20.07	49.0	17	54	−0.358	−2.333	11.590
F	I	secondary	45	28.6	11.99	23.0	19	65	1.408	0.827	1.787
M	I	secondary	24	29.6	13.49	22.0	18	72	1.361	1.499	2.753
F	P	secondary	190	21.0	2.09	20.0	18	34	2.319	9.576	0.151
M	P	secondary	56	21.5	1.51	21.0	19	26	0.463	−0.076	0.202
F	I	higher	93	32.1	9.55	29.0	21	67	1.236	1.155	0.990
M	I	higher	29	37.1	12.43	35.0	22	63	0.474	−1.061	2.307
F	P	higher	28	23.7	2.76	23.0	20	35	2.639	7.780	0.522
M	P	higher	20	25.4	2.41	25.5	22	32	0.702	0.533	0.540

Notes: F: female, M: male, I: Internet subsample, P: paper-and-pencil subsample

Table 2
Self-narrative inclination questionnaire (IAN-R): Conception of sub-scales, and basic psychometric parameters

Self-narration inclination sub-scales	Self-narrative recounting, storytelling (M)	Distancing from experiences (D)	Cultural aspect (K)
Aspects of self narrative activity	communication	identity	culture
Number of items	13	13	4
Number of items with crossloadings	1	5	2
Factor loading:			
M	0.606	0.604	0.624
SD	0.116	0.050	0.040
Min–Max	0.359–0.749	0.527–0.668	0.590–0.673
Common variance	18.0	16.4	9.6
M	45.1	14–65	10.67
SD	9.69	9.56	4.21
Min–Max	13–65	14–65	4–20
Cronbach's α	0.85	0.88	0.75
Inter-scales correlations	M-D $\rho = 0.392$	D-K $\rho = 0.504$	M-K $\rho = 0.517$
Spearman's ρ,	M-K $\rho = 0.517$	M-D $\rho = 0.392$	D-K $\rho = 0.504$

Notes: All results are based on a sample of 519 participants (age: 14 to 72 years, $M = 25.7$, SD = 9.01) consisting of 374 women M age = 25.1, SD = 8.34) and 145 men M age = 27.4, SD = 10.72). Most participants were educated to secondary (64%) or had a higher education qualification (34%). Model from factor analysis ($n = 3$) with principal component analysis with varimax rotation and Keiser normalisation explained 45% of variance (KMO = 0.917, $\chi^2(435) = 5743$, $p < 0.001$).

The quality of the SF was assessed as follows. 1) We compared differences between correlation coefficients with overlapping variance (Steiger, 1980), the proportion of variance the SFs shared with the FF and their classification accuracy. 2) We assessed the reliability and validity of the SFs using the testing group (validation set).

Determining parameters a priori to assess SF quality. The reliability or internal consistency (measured as Cronbach's α of the FF was 0.91 (CI_{95} = [0.89, 0.92]) in the training group and the mean item-item correlation was 0.25. Substituting into the Spearman-Brown formula the minimum acceptable reliability (0.70) and the coefficient of correlation between items it is possible to determine the minimum length of an acceptable SF:

$$r_{tt} = \frac{n \cdot r_{ij}}{1 + (n-1) \cdot r_{ij}} = \frac{n \cdot 0.25}{1 + (n-1) \cdot 0.25} \geq 0.70$$

where n is the length of the FF of the scale, and r_{ij} is the mean correlation between items.

The formula suggests that to have a reliability of at least $\alpha = 0.70$, a SF of the IAN-R should have a least seven items. To ensure that the SF retained the three-factor structure of the original questionnaire (D – distancing from experience; K – the cultural dimension; M –narrative accounting) we applied the above procedure to each subscale (Table 3).

To preserve the reliability of the subscales and ensure that none was reduced to a single item we decided, after analysing the calculations of recommended minimum subscale length (Table 3) we decided to accept an overall length of 10 items and choose 4 items each for scales D and M and 2 for scale K. Moreover, Scale K (4 items only) was abridged beyond the determined length to omit an over-contribution of the items in the whole SF. A consequence of this decision was that the SF subscales would have estimated reliabilities of less than 0.70, but not less than 0.60 (see Table 3). Assuming a 10-item IAN-R SF the theoretical reliability of the whole tool can be calculated using the formula:

$$r_{tt_{SF}} = \frac{\frac{n_{SF}}{n_{FF}} \cdot r_{tt_{FF}}}{1 + (\frac{n_{SF}}{n_{FF}} - 1) \cdot r_{tt_{FF}}}$$

and in our example this amounted to:

$$r_{tt_{SF}} = \frac{\frac{10}{30} \cdot 0.91}{1 + (\frac{10}{30} - 1) \cdot 0.91} \approx 0.771$$

The reduction in reliability coefficient from 0.91 to 0.77 represents a substantial deterioration in reliability and provides one measure of the cost of reducing the length of the IAN-R by two thirds. We concluded that the 14% loss of reliability was compensated for by the likely 1013 minute reduction in completion time (nearly 50%).

After determining an appropriate length for the SF we prepared three SFs based on different statistical techniques. We used R (3.4.3, R Core Team, 2017) and the R-package psych (1.7.8, Revelle, 2017) for all our analyses.

The IRT-based procedure had two phases. In the first stage we selected the items with the highest information value (range: [1.31, 4.78]. Exclusion of below-average values ($M < 2.95$, Md < 2.86 resulted in retention of 19 items, of which 17 were chosen on the basis of the determination coefficient value ($M > 1.25$, Md = 1.26 range= [0.60, 1.75]). Seven of these 17 were subsequently eliminated working by the rule that the retained items should cover the broadest possible range of Θ for the variable measured (detailed IRT results available in the supplementary files and on https://osf.io/5k9cx/).

The RTT-based selection procedure was based on calculations of test reliability values after removal of items: the more removal of a given item reduced reliability the greater the loss in common variance, and hence the more important it was to retain that item in the SF (detailed reliability results available in the supplementary files and at https://osf.io/5k9cx/). In other words, the procedure selected those items whose removal had most impact on scale reliability. Another coefficient of scale reliability was proposed by Hayes (2005), who suggested that all SFs containing from 2 to $k - 1$ items should be constructed and used to calculate average values of Cronbach's α and the coefficient of correlation between the SFs and FF. Hayes argued that only these averages should be used for item selection (which should be based on maximization of Cronbach's α and the SF-FF correlation coefficient). According to Hayes, relying on the hierarchical 'leave-one-out' procedure results in failure to consider many possible item combinations. This procedure offered by most statistical packages removes the weakest item (the one which has least effect on scale reliability) and then re-calculates the reliability of the remaining items. The number of possible items combinations is given by the following formula: $50(k^2 + k - 6)/(2^k - k - 2)$, where k is the number of items, and may lead to inappropriate selection of items for SFs. However, the opposite approach choosing the strongest items does not suffer from this problem.

The set of the items selected using Hayes's procedure differed from the set created according to our criteria with respect to just one item (#2 instead of #6). Because our RTT version and Hayes's version do not differ in correlation coefficients from SFs created using other methods and the full version, we have opted to present a version which was easier to prepare (most researchers construct SFs based on reliability indices available in popular statistical packages and Hayes's procedure is time-consuming).

The third SF was based on factor loadings obtained from EFA with varimax rotation. We selected the items which had the highest factor loadings for each subscale (detailed EFA results available at https://osf.io/5k9cx/).

Table 3
Reliability of IAN-R subscales and the number of items for SF scale

Subscale	α	λ_6	Average r	M	SD	Calculated N	Accepted N	r_{ttSF}
D [13]	0.876	0.882	0.354	3.821	0.732	4.27	4	0.685
M [13]	0.872	0.885	0.347	3.462	0.756	4.38	4	0.677
K [4]	0.754	0.702	0.434	2.658	1.058	3.05	2	0.605

Note. In square parenthes are number of items. Number of paricipants = 480. For calculated N assumed $r_{\text{ttSF}} = 0.70$.

Comparing SFs and Evaluating Abridging Techniques. The IAN-R items making up the three SFs are presented in Table 4. The overlap in selection of items measured as Fleiss's κ ($\kappa = 0.30$; $p < 0.01$) was low. In other words the composition of the SF varied considerably according to the technique used to derive it. Only three items (#5, #27 and #29) were included in all SFs.

After preparing various SF of the IAN-R questionnaire (on the training data - 80% of the original sample) we proceeded to assess the quality of each version based on data from the testing group (20% of the original sample). We started by calculating mean scores for the FF and SFs. Using the Friedman test with Holm's correction for the comparison of pairs, we found differences between the versions in the cases of scales D and K (scale D: $\chi^2(3, N = 96) = 65.99$; $p < 0.001$; for Scale K: $\chi^2(3, N = 100) = 27.41$; $p < 0.001$. No differences in the results for Scale M were found: $\chi^2(3, N = 100) = 7.18$; $p = 0.066$.

In the case of scale D the EFA- and RTT-based SFs resulted in lower means ($p < 0.001$) than the FF, but the IRT-based SF produced a similar mean D score ($p = 0.116$). The mean K scores for the IRT- and RTT-based SFs were higher in absolute terms than the mean K score for the FF, but the difference was not significant ($p = 1$). The mean K score for the EFA-based version was lower than those of the IRT- and RTT-based versions ($p < 0.001$). Scale M had the least varied results – here only in comparison with the full scale score, the IRT version results were significantly overvalued ($p < 0.01$. The confidence interval of average scores for particular versions is presented in Figure 1.

Having summed up the differences between the various SF, we suspect that the size and direction of the differences depend on the content of SF items. None of the techniques provided the results which were identical to the full scale. Scale D was most sensitive to abridgement – only the IRT-based version delivered similar results to the FF. In the case of scale K the EFA-based SF produced different results from the FF, as did the IRT-based SF in the case of scale M. We tentatively concluded that the IRT- and RTT-based SF are closer to the FF than the EFA-based SF.

We also calculated coefficients of correlation between SF scores and FF scores and then assessed pairwise differences between these coefficients (Table 5). The method we used takes into account the variance common to pairs of correlation coefficients (Steiger, 1980).

In the case of the D scale the IRT-based SF produced scores that were more highly correlated with those of the FF than did the EFA-based SF, but they were not more highly correlated than D scores on the RTT-based SF. In the case of the M scale the IRT-based SF scores were more highly correlated with the FF score than were those of the RTT-based SF ($\alpha = 0.10$). In the case of scale K all the SFs were similarly correlated with the FF, although in absolute terms the correlation was highest in the case of the IRT-based SF.

3.4 Evaluation and theoretical reliability

We selected the IRT-based SF as the final SF and calculated reliability coefficients for the testing group (Table 6).

The overlap in variance between the final SF and FF ($r_{ttSF} \cdot r_{ttFF}$) was greatest in the case of scale D (65.7%) and least in the case of the shortest scale, scale K (36.2%). In the case of scale M the percentage of shared variance was 54.9%. Given that the SF is 67% shorter than FF, the scale K result is not surprising, and SFS scales D and K show moderate overlap in variance with FFs of the scales.

Content validity assessed on testing sample by CFA method revealed good fit coefficients (CFI = 0.98, TLI = 0.97, AGFI = 0.97, RMSEA = 0.07, $\text{CI}_{95} = [0.038, 0.099]$, $p = 0.14$. These values apply to a model in which correlations between items from different subscale were allowed (#5 and #7; #5 and #6; #21 and #29; #21 and #7), indicating an increase in the general IAN-R factor and a certain loss of content specific for particular subscale.

The final indicator we used to assess the quality of the new short form was classification accuracy. Participants in the testing group were classified into low- and high-narrative subgroups based on whether their score was above or below the group mean score, using both the FF and SF. We then compared the classifications (Table 7).

The SF correctly classifies 85.1% of cases. It performed better with low-narrative cases (1 out of 20 wrongly classified) than with high-narrative cases (13 out of 74 wrongly

Table 4
Items that were selected using IRT, RTT and EFA

Method	Scale D				Scale K		Scale M			
IRT	6	17	22	29	5	18	7	21	27	30
RTT	2	22	24	29	5	9	19	20	21	27
EFA	2	8	24	29	5	13	19	20	27	30

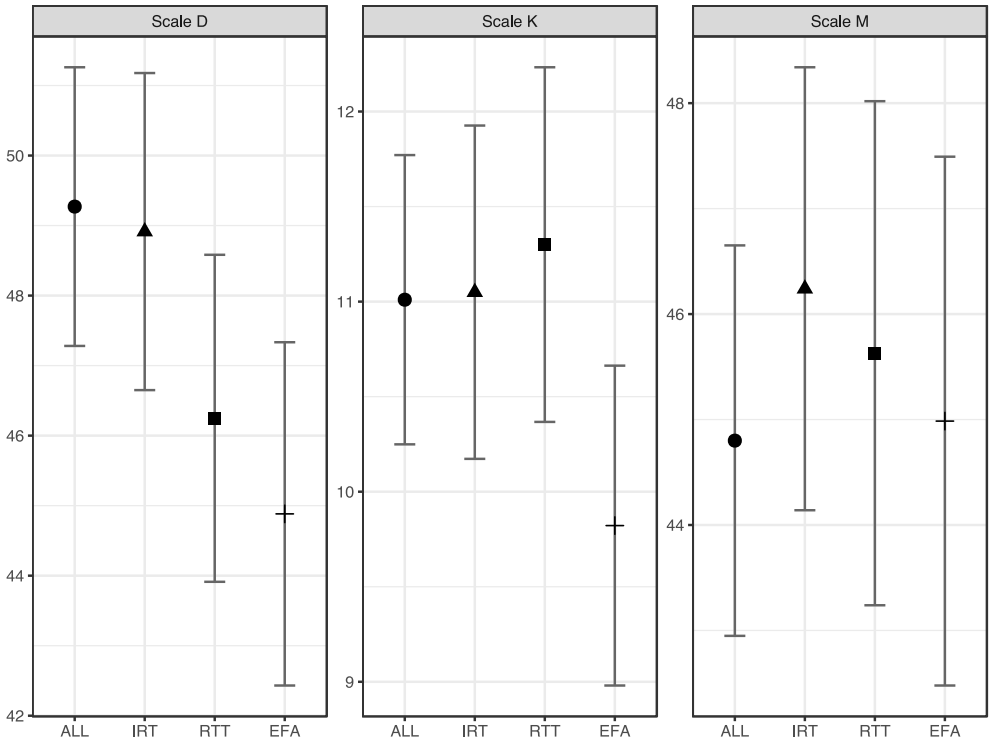

Figure 1. Means with a 95% confidence interval for scale scores for the orginal questionnaire (FF) and the IRT-, RTT-, EFA-based short forms.

classified). In other words the SF produces a more conservative assessment of narrative inclination than the FF; this is clearly seen in Fig. 2, which shows density distributions for scores on both forms.

Taking into account the results of all the analyses the final 10-item SF of the IAN-R can be recommended. The results obtained using the SF were highly correlated with the results obtained using the FF. Our analyses also suggest that IRT-based selection of items has little advantage over the commonly used CTT-based methods, but we cannot be confident of this conclusion until it has been confirmed by analysis of other tools (see also Kleka, 2013).

4 Discussion

A research tool which provides a reliable, valid measurement of mental properties, yet is not onerous for subject remains the holy grail of psychometric researchers. Within psychological research there is a move towards use of shorter tools, driven partly by the demand for the research into psychological constructs in other fields (see also Ziegler, Kemper, & Kruyen, 2014). Another factor in the preference for short tools is that it is very common to administer a battery of tests and subjects are increasingly reluctant to submit to lengthy test session. The paradoxically higher face validity of short tools is also relevant (Wanous, Reichers, and Hudy, 1997, cf. Rammstedt and Beierlein, 2014. The procedure we have proposed could be used to bring a degree of order to the abridgement process, by creating an easily attainable standard which ensures that the psychometric history of a tool is transparent and an informed decision about its application. The preliminary reflection on abridgement that we suggest should eliminate ill-advised attempts to abridge tools which are of poor quality in the first place.

The recommendation for describing the statistical technique selected to abridge a given questionnaire will enable meta-data collection and allow tracing a relation between a

Table 5
Coefficients of correlation between SF scale scores and FF scores and the significance of differences between them. The pairwise comparisons of correlations take into account common variance using Steiger's method

		Corr. with full form			Significance of differences					
		IRT	RTT	EFA	IRT–RTT		IRT–EFA		RTT–EFA	
Subscale	n	*r*	*r*	*r*	*z*	*p*	*z*	*p*	*z*	*p*
M	100	0.909	0.869	0.820	1.86	0.066	3.62	< 0.001	2.70	0.008
D	96	0.937	0.921	0.879	1.12	0.260	3.04	0.003	2.92	0.004
K	100	0.891	0.864	0.835	1.01	0.320	1.92	0.058	0.88	0.380

Table 6
Empirical and theoretical reliability

		CI_{95}		Average				theoretical	
Subscale	α	lower	upper	λ_6	*r*	*M*	SD	r_{ttSF}	r_{ttFF}
M [4]	0.63	0.44	0.81	0.57	0.30	3.6	0.82	0.677	0.872
D [4]	0.75	0.60	0.90	0.72	0.43	3.7	0.88	0.685	0.876
K [2]	0.48	0.14	0.82	0.32	0.32	2.8	1.10	0.605	0.754

technique and obtained set of items. For the time being we do not have access to such data and creating an appropriate database would require the cooperation of several researchers or research teams.

From a researcher's point of view it is desirable to know biases of statistical techniques (e.g. favouring items from the core component of a target construct or cultural biases) used for developing SF and adapting it for specific research purposes. Relying solely on reliability statistics, regression or factor analysis (which is very easy thanks to the modern statistical packages) produces variable results and there is a need for theoretical reflection before the calculation stage. Although automated iterative automated methods for abridging questionnaires have produced promising results they are not without problems. Using a GA-based or ACO-based approach without taking into account the psychometric properties of the original tool could result in obtaining tools with unknown properties. Moreover, there is a risk that they might gain some acceptance simply because they had been developed using a computationally sophisticated but not necessarily psychometrically sound techniques. Requiring reports on newly developed tests to include an assessment of criterion validity would help to prevent the dissemination of low-quality tools, but it would also impose a requirement for much deeper methodological reflection and additional testing, and history shows that this is too much to expect.

Our analyses suggest that SFs produced using different statistical techniques yield rather similar scores in practice (at least with regard to the constructs analysed so far; cf. Kleka, 2013; Kleka and Paluchowski, 2017, but the most advanced techniques, which apply the most stringent assumptions, i.e. IRT models, offer greatest control over how the abridgement affects the psychological meaning of scores on an instrument. Generalisability theory (GT)-based approaches to abridgement also seem promising (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Generalisability theory offers detailed insight into particular sources of measurement error and it allows one to divide CTT's error into components based on their source: 1) test items, 2) subjects, 3) measurement time (test-retest) and 4) interactions between these sources (Ziegler, Poropat, & Mell, 2014).

The difficulty of meeting IRT assumptions and the low popularity of GT mean that it may be acceptable to base a SF on reliability analysis, but it should be remembered that relying sole on analysis of Cronbach's α will result in loss of content validity in the SF (Ziegler, Poropat, & Mell, 2014). It would therefore be better to base selection decisions on analysis of McDonald's ω, which is not as strongly influence by test length and number of observations as Cronbach's α and is robust against frequent violation of assumptions regarding tau-equivalent items.

5 Proposal for a Standard Abridgement Procedure

The researcher intending to develop a SF of an existing tool should be aware that there are many steps before one can select items for the SF and that the process does not end when a SF has been constructed. Preparatory activities and analyses, which frequently include complex theoretical analysis of the target construct (cf. Ziegler, Kemper, & Kruyen, 2014), and empirical verification of the properties of the new tool are very important parts of the abridgement process. Based

Table 7
Classification of the subjects from validation subsample according to SF and FF according mediana [M]

Score	IAN-$R_{FF} < M$	IAN-$R_{FF} \geq M$	Total
IAN-$R_{SF} < M$	19	13	32
IAN-$R_{SF} \geq M$	1	61	62
Total	20	74	94

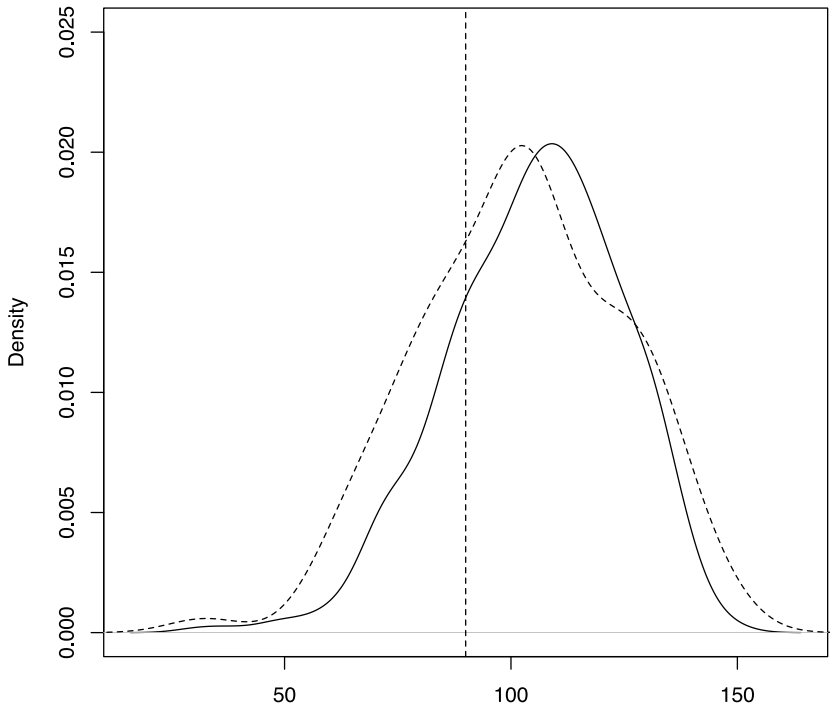

Figure 2. Density distributions on the FF(solid line) and SF (dashed line) of the IAN-R.

on our review of the literature and the research described here we recommend that psychologists seeking to abridge a tool carry out the following sequence of steps.

1. Characterise the desirable properties of the SF (psychometric properties, purpose and factorial structure).

This stage enables verifying the quality of the FF (questionnaire with low validity and reliability should not be abridged) and defining the goal of constructing a new version. Is the new SF intended to 1) characterise the subject with respect to a continuous trait or 2) determine the group into which the subject falls with respect to a target variable? At this stage, a researcher should answer a question whether the goal of the tool remains this same. If any change has been assumed, then one needs to think which statistical technique allows a better selection of items for the new objective. At this point, it needs to be emphasised that changing the objective potentially entails a new definition of validity, which will affect the comparison of SF and FF scores.

Knowledge about tool structure is required to plan appropriate analyses. In the case of multifactorial tools, comparisons of factor scores as well as overall scores will be required and there is the additional problem of preserving the pattern of relationships between the factors when the tool is abridged.

2. Analyse potential costs and benefits of abridgement in your particular case.

This should be done not only with reference to psychometric properties themselves, but should also take into account the broader context of the research (what it implies, etc.). Researchers should estimate the potential savings of time and other resources, the potential loss of both reliability and correlation with the criterion. At this stage a decision about the length of the SF should be taken.

The practical factors which are worth considering include changes in the psychometric properties of the tool as a result of abridgement, changes to the method of testing, changes to scores and what limits there should be on application of the new tool. For instance, Silverstein (1990) indicated that SFs of the WAIS should be used with caution. It is also advisable to investigate the "justification-of-use" (Lee et al., 2014, p. 518) and, especially, the practical consequences of employing SFs in particular circumstances, possible changes to the human and financial resources required and the impact on research policies (multi-level circumstances). Good decisions about whether and how to proceed with abridgement depend on precise analysis of empirical and theoretical criteria.

3. Choose the statistical technique that will be used to select items.

At this stage one can decide which of numerous statistical techniques to employ, based on one's knowledge of the structure and content of the original. We recommend IRT-based techniques wherever the assumptions can be met, but our analyses for this study suggest that abridgement based on reliability coefficients produces SFs that are almost as good and it is easier to perform.

4. Select item for the SF.

Relying solely on statistical analyses would be risky and so we recommend taking into account the structure of the SF, the use to which it will be put (cf. Point 1), and the content of the items. The content of a given item may be so important that it is worth including it in a SF, despite the results of statistical analyses, which may be biased. Even when the measured theoretical construct does not suffer because of questionnaire abridgement, the length (a number of items) of the SF has a big influence on its reliability and classification accuracy. Failure to include sufficient items that are capable of discriminating between subjects at one extreme of the distribution of the trait being measured will lead to a skewed distribution of scores and make the SF vulnerable to floor or ceiling effects (Anastasi & Urbina, 1989).

5. Describe the psychometric parameters of the SF.

The ideal would be to conduct a pilot study with the new SF and determine its psychometric properties this way. If this is not possible information about the psychometric properties of the SF should be obtained in another way. We suggest splitting the sample recruited for development of the SF into two sets and using one set to construct the SF and the other set to calculate reliability and validity parameters.

6. Evaluate the SF and confirm its psychometric properties.

Evaluation of the SF requires constant monitoring of its various parameters. This is a process encompassing proper preparation for abridgement, abridgement itself, and evaluation of the shortened tool. The last stage consists of comparing the properties assumed a priori with the properties obtained e.g. in the pilot study, and if necessary also considering the various contexts in which the tool might be used (see also Cramer, Wevodau, Gardner, & Bryson, 2017). If the scores obtained in the pilot study do not reach the target reliability and validity level, one should return to Point 4 and re-select items or return to Point 2 and verify the estimate of the acceptable length of the SF and re-evaluate the goal of abridgement.

Evaluation criteria for SFs mainly deal with proper representation of the measured construct, or a uniform allocation of items in subscales if the tool is multi-dimensional. These problems may be described by empirical criteria, the most important of which are: 1) SF reliability; 2) validity (mainly face and content validity relative to the FF); 3) the extent to which the SF reproduces the structure of the FF (this is the criterion is often used in abridgement reports); 4) similar pattern of relationships with other variables to the FF, particularly with respect to demographic variables, such as age or education, but also personal traits and characteristics (see also Lee et al., 2014).

The last stage of SF creation should consist of assessing the classification accuracy of the SF by comparing the results with classification based on the FF. Lee et al. (2014) suggested a threshold of $r > 0.90$ of correlation between raw scores on the SF and FF. Ultimately the SF cannot be significantly different from FF nor should it serve another goal than the one selected by a researcher.

6 Limitations and Directions for Further Research

There are some potential limitations of the results presented here linked to the lack of systematic studies on statistical techniques in relation to different types of tools. So far we have examined several psychological questionnaires, with similar results: various statistical techniques recommend clearly different items for SFs. It seems that the final score is more dependent on the content of items than their statistical properties (cf. Kleka, 2013; Kleka & Paluchowski, 2017). We carried out analyses not reported in this article,

examining constructs from various psychological domains (intelligence; temperament; workaholism; self-narrative inclination) that are measured with multi-dimensional tools based on Likert scales. Research tools with a different format (e.g. Q-sort) cannot be abridged using the procedure we have proposed here.

The procedure described here for abridging a tool is based primarily on conventional psychometric criteria such as internal consistency and factor structure rather than on pragmatic considerations. Promising results have been achieved using modern computing tools and sophisticated computations to abridge psychological instruments. The generally limited programming skills of social science researchers remain a significant barrier to adoption of such techniques. At present effective application of e.g. the GA method or ACO method requires the ability to program in R, which is not common amongst social science researchers. But the barriers to use of computationally sophisticated techniques will continue to diminish, with the spread of modern software that draws on techniques applied in other disciplines. The danger of this going down this path is that it leads to widespread use of a "blackbox" approach to abridgement (using a method that appears to work, although one does not understand how), which could result in tools being developed without sufficient reflection on their accuracy, reliability and applicability.

References

Anastasi, A. & Urbina, S. (1989). *Psychological testing*. London: Collier Macmillan.

Arthur, W. & Day, D. V. (1994). Development of a short form for the raven advanced progressive matrices test. *Educational and Psychological measurement*, *54*(2), 394–403.

Bartholomew, D. J. (1987). *Latent variable models and factors analysis*. London: Oxford University Press, Inc.

Bersoff, D. N. (1971). Short forms of individual intelligence tests for children: Review and critique. *Journal of School Psychology*, *9*(3), 310–320. doi:10.1016/0022-4405(71)90088-4

Bors, D. A. & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, *58*(3), 382–398. doi:10.1177/0013164498058003002

Choynowski, M. (1971). Podstawy i zastosowanie teorii rzetelności testów psychologicznych [Fundamentals and applications of the theory of the reliability of psychological tests]. In J. Kozielecki (Ed.), *Problemy psychologii matematycznej* (pp. 65–118). Warszawa: PWN.

Clara, I. P. & Huynh, C.-L. (2003). Four short-form linear equation estimates of Wechsler Adult Intelligence Scale III IQs in an elderly sample. *Measurement and Evaluation in Counseling and Development*, *35*(4), 251–262.

Coroiu, A., Meyer, A., Gomez-Garibello, C. A., Brähler, E., Hessel, A., & Körner, A. (2015). Brief form of the interpersonal competence questionnaire (ICQ-15): Development and preliminary validation with a German population sample. *European Journal of Psychological Assessment*, *31*(4), 272–279. doi:10.1027/1015-5759/a000234

Coste, J., Guillemin, F., Pouchot, J., & Fermanian, J. (1997). Methodological approaches to shortening composite measurement scales. *Journal of Clinical Epidemiology*, *50*(3), 247–252.

Cramer, R. J., Wevodau, A. L., Gardner, B. O., & Bryson, C. N. (2017). A validation study of the need for affect questionnaire– short Form in legal contexts. *Journal of Personality Assessment*, *99*(1), 67–77. doi:10.1080/00223891.2016.1205076

Crawford, J. R., Allan, K. M., & Jack, A. M. (1992). Short-forms of the UK WAIS–R: Regression equations and their predictive validity in a general population sample. *British Journal of Clinical Psychology*, *31*(2), 191–202.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley: John & Sons.

Fox, R. A., McManus, I. C., & Winder, B. C. (2001). The shortened Study Process Questionnaire: An investigation of its structure and longitudinal stability using confirmatory factor analysis. *British Journal of Educational Psychology*, *71*(4), 511–530. doi:10.1348/000709901158659

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282.

Hamel, R. & Schmittmann, V. D. (2006). The 20-minute version as a predictor of the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement*, *66*(6), 1039–1046.

Hayes, A. F. (2005). *A computational tool for survey shortening applicable to composite attitude, opinion, and personality measurement scales*. Paper presented at the meeting of the Midwestern Association for Public Opinion Research in Chicago.

Kaufman, A. S. (1972). A short form of the Wechsler Preschool and Primary Scale of Intelligence. *Journal of Consulting and Clinical Psychology*, *39*(3), 361–369.

Kerlinger, E. N. & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston.

Kleka, P. (2013). *Zastosowanie teorii odpowiadania na pozycje testowe (IRT) do tworzenia skróconych wer-*

sji testów i kwestionariuszy psychologicznych [Use of the Item Response Theory (IRT) to produce abridged versions of psychological tests and questionnaires] (Doctoral dissertation, Adam Mickiewicz University in Poznan). Retrieved from http://hdl.handle.net/10593/5943

Kleka, P. & Paluchowski, W. J. (2017). Shortening of psychological tests–assumptions, methods and doubts. *Polish Psychological Bulletin*, *48*(4), 516–522.

Konarski, R. (2014). *Modele równań strukturalnych: Teoria i praktyka [Structural equation modeling: Theory and practice]*. Wydawnictwo Naukowe PWN.

Las Hayas, C., Quintana, J. M., Padierna, J. A., Bilbao, A., & Muñoz, P. (2010). Use of rasch methodology to develop a short version of the Health Related Quality of life for Eating Disorders questionnaire: A prospective study. *Health and Quality of life Outcomes*, *29*(8), 1–12.

Lee, E. S., Bygrave, C., Mahar, J., & Garg, N. (2014). Can higher education exams be shortened? A proposed methodology. *International Journal of Information and Education Technology*, *4*(6), 517–525. doi:10.7763/IJIET.2014.V4.462

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*(3), 411–431.

Manos, R. C., Kanter, J. W., & Luo, W. (2011). The behavioral activation for depression scale–short form: Development and validation. *Behavior Therapy*, *42*(4), 726–739.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and Quality of Life Outcomes*, *27*(1), 1–13.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rammstedt, B. & Beierlein, C. (2014). Can't we make it any shorter? The limits of personality assessment and ways to overcome them. *Journal of Individual Differences*, *25*(4), 212–220. doi:10.1027/1614-0001/a000141

Revelle, W. (2017). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Richmond, VA: Psychometric Society.

Satz, P. & Mogel, S. (1962). An abbreviation of the WAIS for clinical use. *Journal of Clinical Psychology*, *19*(3), 1877–1879.

Schaufeli, W. B., Bakker, A. B., & Salanova, M. (2006). The measurement of work engagement with a short questionnaire: A cross-national study. *Educational and Psychological Measurement*, *66*(4), 701–716.

Schipolowski, S., Schroeders, U., & Wilhelm, O. (2014). Pitfalls and challenges in constructing short forms of cognitive ability measures. *Journal of Individual Differences*, *35*(4), 190–200. doi:10.1027/1614-0001/a000134

Silverstein, A. B. (1990). Short forms of individual intelligence tests. *Psychological Assessment*, *2*(1), 3–11. doi:10.1037/1040-3590.2.1.3

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*(1), 102–111. doi:10.1037//1040-3590.12.1.102

Soroko, E. (2013). Kwestionariusz Inklinacji Autonarracyjnej (IAN-R)–pomiar skłonności do narracyjnego opracowywania i relacjonowania doświadczenia [Questionnaire on auto-narrative inclination (IAN-R) - measurement of susceptibility to narrative development and experience reporting]. *Studia Psychologiczne*, *51*(1), 5–18. doi:10.2478/v10167-010-0063-4

Soroko, E. (2014). *Aktywność autonarracyjna osób z różnym poziomem organizacji osobowości. Opowieści o bliskich związkach [Self-narrative activity of people with different levels of personality organisation. Stories about close relationships]*. Poznań: Wydawnictwo Naukowe UAM.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251.

Tabachnick, B. G. & Fidell, L. S. (2001). *Using multivariate analysis*. Boston: llyn & Bacon/Pearson Education.

Takane, Y. & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.

Van der Elst, W., Ouwehand, C., van Rijn, P., Lee, N., Van Boxtel, M., & Jolles, J. (2013). The shortened raven standard progressive matrices: Item response theory–based psychometric analyses and normative data. *Assessment*, *20*(1), 48–59.

van Dierendonck, D., Díaz, D., Rodríguez-Carvajal, R., Blanco, A., & Jiménez, M. (2008). Ryff's six-factor model of psychological well-being, a spanish exploration. *Social Indicators Research*, *87*(3), 473–479.

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: how good are single-item measures? *Journal of Applied Psychology*, *82*(2), 247–252.

Warrington, E. K., James, M., & Maciejewski, C. (1986). The WAIS as a lateralizing and localizing diagnostic instrument: A study of 656 patients with unilateral cerebral lesions. *Neuropsychologia*, *24*(2), 223–239.

Woo-Sam, J. & Zimmerman, I. L. (1973). Note on applicability of the Kaufman formula for abbreviating the WPPSI. *Perceptual and Motor Skills*, *36*(3_suppl), 1121–1122.

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, *44*(2), 180–198. doi:10.1016/j.jrp.2010.01.002

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales – five misunderstandings and ways to overcome them. *Journal of Individual Differences*, *35*(4), 185–189. doi:10.1027/1614-0001/a000148

Ziegler, M., Poropat, A., & Mell, J. (2014). Does the length of a questionnaire matter? Expected and unexpected answers from generalizability theory. *Journal of Individual Differences*, *35*(4), 250–261. doi:10.1027/1614-0001/a000147