Other Papers

Polish Psychological Bulletin
2017, vol. 48(4) 516–522
DOI - 10.1515/ppb-2017-0058

*Paweł Kleka**
*Władysław Jacek Paluchowski**

Shortening of psychological tests – assumptions, methods and doubts

Abstract: *In this article, on the basis of questionnaire data collected for other purposes, the Authors want to show the consequences of various methods of shortening of tests and what may result from such an action for diagnosticians, researchers and examined individuals. The research aim of the work is to show the best method of shortening of the scale of questionnaires. Will shortening of a questionnaire according to different statistical techniques bring the same results? Will the quality of shortened scales be comparable? Is any of statistical techniques better for shortening of the scale of a questionnaire? The obtained results suggest a poorly controlable effect of the methods of questionnaire shortening. Moreover creating a short version on the basis of the results collected with the use of the full version leads to obtaining a tool with unknown diagnostic and psychometric properties.*

Keywords: *reliability, validity, shortening tests, abbreviated version*

Actions for shortening of psychological tests

An obvious purpose of shortening of psychological tests is the desire to shorten test duration. Shortening of psychological tests concerned all methods: projective (TAT and CAT – Bellak, 1955; Chushmir, 1985; TEMAS – instead of the full 23-card version – a shortened version of 9 cards – Costantino, Malgady, Vazquez, 1981), questionnaire (MMPI, MMPI-2) or techniques of examining intellectual ability (WAIS, WISC). Let us take a look at two selected examples.

The MMPI questionnaire is a particularly long method: it consists of 566 items resulting in an obvious tendency to shorten to 168 and even 71 items. What is easy in the case of single-scale questionnaires (reducing the number of items), becomes a serious problem in the case of multi-scale questionnaires. Such a questionnaire becomes then a real challenge: in its classic version (MMPI) has 3 validity scales and 10 clinical scales, in the current version (MMPI-2) it has 2 extra validity scales and 15 extra content scales. In 1946–1974, 13 various short versions of this questionnaire were created (Stevens, Reilley, 1980). The best-known short versions are: MMPI-168 (Overall, Gomez-Mont, 1974), Faschingbauer's FAM version (Faschingbauer, 1974) and Mini-Mult (Kincannon, 1968). Kincannon's Mini-Mult comprises 71 items, chosen due to the representativeness of their contents for MMPI control and clinical scales. Faschingbauer (1974), as part of his doctoral thesis, created a version of MMPI referring to the grouping procedure in such a way so that the correlation between the full and short version of chosen scales (F, Pp, Mf, Pa, Sc, Hy, Si) would not be smaller than .85; the remaining questions came from the scales of the Mini-Mult version. Sometimes a procedure of creating a shortened version was based on a completely incidental criterion. The MMPI-168 version (Overall, Gomez-Mont, 1974) was created from the first 168 items of the MMPI questionnaire. Also Dahlstrom and Archer (2000) created a shorter version of MMPI-2 simply selecting the first 180 items (MMPI-180). Unfortunately, as it turned out later, short versions have flaws regarding both their low reliability and the loss of accuracy (Butcher, Hostetler, 1990; Butcher, Kendall, Hoffman, 1980).

Also, with regard to the tests examining intellectual ability an argument of necessary shortening of test duration was raised. Like in the case of questionnaires, shortening consisted in the reduction of items in a scales. Also in the case of testing intelligence a basic practice, besides the above mentioned, was reducing the subtests of a battery to test intellectual ability (Groth-Marnat, 2003, pp. 191–195)[1].

[1] In the case of the MMPI questionnaire, there exists no general score (e.g. the sum of the scores of clinical scales) which would have a diagnostic

* **Uniwersytet im. A. Mickiewicza, Instytut Psychologii**

Corresponding author: Paweł Kleka, e-mail: kleka@amu.edu.pl

The most common technique of testing intelligence was and is one of the forms of batteries of Wechsler's tests, enabling the calculation of verbal, non-verbal and full-scale intelligence quotient. Philip Levy (1968) described five different methods of shortening of Wechsler's batteries. Generally, these strategies are either a selection of tests (*scale sampling*) or a selection of tasks from subtests (*item sampling*). One of them was a choice (depending on needs) of subtests of the verbal (verbal intelligence quotient) or non-verbal scale; another a selection of subtests to estimate the score in the full scale. The criterion was a connection of a subtest with a given intelligence quotient. An example of such conduct is Doppelt's version (1956), consisting of the subtests: Vocabulary, Arithmetic, Block Design and Picture Arrangement, correlating with the full scale (depending on a group) from .93–.95. Corresponding subtests were chosen by Silverstein (1982a, 1990). A slightly different approach was adopted by Jones (1962) who searched for the best set of subtests with the use of a multiple regression equation. On the other hand, Ward, Selby and Clark (1987) selected subtests being guided both by the criterion of correlation and the shortest duration of a test (cf. implementing this criterion by Kaufman et al., 1991). Yet another method of selecting subscales is a reference to the results of the factor analysis. The original version of the battery WAIS-III was replaced by The Psychological Corporation[2] in 1997 with the Wechsler Abbreviated Scale of Intelligence battery (WASI). It consisted of four subscales (Vocabulary, Similarities, Block Design and Matrix Reasoning), which was supposed to enable accurate estimation of the verbal and non-verbal scale and the full, original battery. The criterion for choosing these subtests was their high factor loading of the factor *g* (cf. the description of the structure of loadings Canivez et al., 2009). Another short version is a version comprising two subtests: Vocabulary and Block Design; the scales were chosen on the basis of an analogical criterion. A similar strategy was employed also for the WAIS-III version by Ward and Ryan (1999) who created a version consisting of seven subtests (Information, Digit Span, Arithmetic, Similarities, Picture Arrangement, Block Design and Digit Symbols).

Another approach is the selection of items (tasks) in subtests and, thus, shortening of Wechsler's battery and test duration. The most radical way (cf. Silverstein, 1982b) was used by Satz and Mogel (1962) who chose tasks from the WAIS battery (and then from WAIS-III and WISC-III) according to a formal criterion. For the scales: Information, Vocabulary and Picture Arrangement it was every third task, and for the rest: the tasks having odd numbers (similar to the method of calculating test reliability).

Analysing the issues of test shortening from a technical point of view, it can be said that authors of various undertakings who chose elements (items or scales) for a short-form scale or battery refer either to the correlation of selected items with the general score, to the criterion of representativeness of the content of a short test (mainly the factor analysis, and also grouping of items based on their content), or to a diagnostic value of selected elements (i.e. an increase in accuracy of a classifying decision, using e.g. an analogy to adaptive testing) or finally to any formal criteria (e.g. every *n* item, the first *n* items, etc.).

Smith, McCarthy and Anderson (2000), describes fundamental mistakes made while shortening of psychological tests. Most of all, the biggest reservation about short forms is their lack of accuracy in relation to the original, which clearly does not offset the benefits concerning test duration. However, to what extent a short version is inferior to the original also depends on the methodology of test shortening and this is the main focus of the authors. For instance, they stress that relying statistical decisions concerning the shortened scales on the data collected with the use of original tools may lead to errors, especially pertaining to the correlation of the short version and the original.

One of the most basic faults of test shortening, as claimed by Smith et al. (2000), is the assumption that the short version will have the same psychometric properties as the original scale. However, very frequently the theoretical accuracy of the shorter form is lower than the original one and only referring to the latter will not suffice. The major proof of a lowered accuracy is the fact that the short scale covers only a fragment of the area of observational identification of the examined construct. Because it is a kind of idealisation to assume that every test item concerns precisely the same aspect of the construct which is of interest to us. If it were like this, the researcher would not have to check the accuracy of a new tool. Moreover, choosing those items of the test which correlate best with the general score of the scale (which seems to be a reasonable criterion of selecting items for the short version) in the case of scales of a relatively low homogeneity, one would leave out certain areas of the empirical indicator of the examined phenomenon in an uncontrolled way. Sometimes it may be crucial for the accuracy of decisions made on the basis of the score of the shorter version. Therefore, Smith et al. (2000) suggest that test shortening be supplemented with the analysis of its content to guarantee its representative share in the short version or the factor analysis of test results be carried out. Furthermore, researchers inaccurately assume that, because the scale is shorter, lower values of psychometric properties are acceptable. Obviously, as shown by Smith et al. (2000), the standards for short versions of the tool should be the same as for the full, long version.

Research problem

The main research problem is comparing various statistical techniques on the basis of which researchers take decisions about including or excluding an items. When one tries to estimate the relations between statistical techniques, a question arises how to assess the equivalence

sense. It is different in the case of intelligence testing techniques where such a general score is IQ (general, verbal or non-verbal).

[2] Professional organisation founded in 1921 in the state of New York, USA.

of obtained results. For the purposes of analyses, in the present work the Authors assumed an external criterion: a diagnosis according to ICD-10, which enabled an objective comparison of the quality of short scales and selection of the best statistical technique for shortening of a questionnaire. Having an external criterion, the Authors were able to rely the evaluation of quality on the area under an Receiver Operating Characteristic curve (ROC); higher values point to a better diagnostic power manifesting itself through a higher sensitivity and specificity of the tool.

The review of various research studies does not point clearly to any of the statistical techniques as the best method of shortening, but even a cursory review of the literature indicates a big proportion of the factor analysis and reliability analysis. The reason is certainly the presence of these techniques in popular statistical software packages and easy interpretation of results, even without a deeper knowledge about the assumptions of the techniques. Other techniques chosen for the analysis on which a researcher can rely their decisions about usefulness of particular items are logistic regression (requiring an external criterion) and item analysis with the use of the IRT paradigm.

The chosen area of checking the usefulness of statistical techniques selected for test shortening will be the normalisation scores of the Working Excessively Questionnaire (WEQ), and more specifically one of its scales: Lack of Control over Work Scale (LCWS) (Paluchowski, Hornowska, Haładziński, Kaczmarek, 2014). The scale has been chosen because of the closest relation to the external criterion: a diagnosis independent from scale scores.

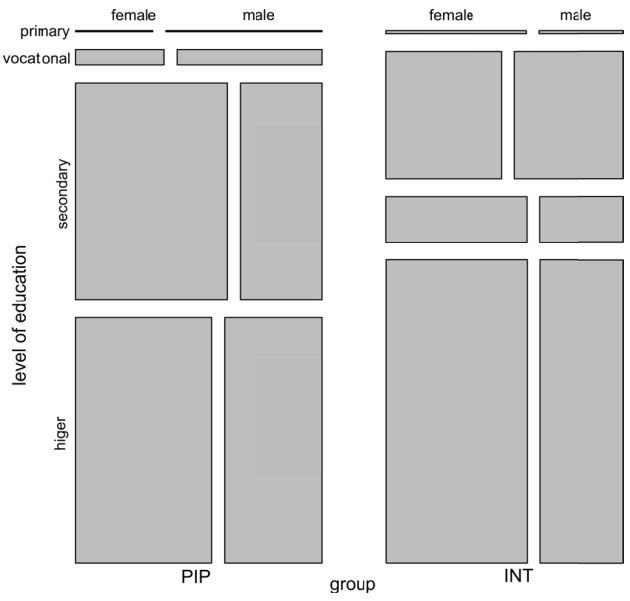
Method

Participants

The study analysed the results (2658 records) gathered during normalisation research. The data were collected with the use of the traditional paper-and-pencil method ($N_{PIP} = 1388$, 52%) and on the Internet ($N_{INT} = 1270$, 48%). The people from the PIP group were on average 4.6 years older than the people from the INT group whose average age amounted to 30.6 years ($t(2504) = 12.96$, $p < .001$, *95%CI* = 3.96–5.37). In the tested sample women were the majority (60%) and it applied to both subgroups. However, sex did not differentiate the age of the examined individuals ($t(2142) = 1.27$, $p = .175$). Secondary and higher education prevailed in the PIP group and vocational and higher education in the INT group. Detailed proportions of the variables in the tested group are presented in the Figure 1.

Subject reported the total length of work experience and the length of service in the current place of work. It amounted to, respectively, 10 years ($SD = 9$) and 5.6 years ($SD = 6.5$). The longest total work experience amounted to 52 years (43 in the current place of work), and the shortest 3 years (one year in the current place of work). No differences between men and women were noted in the total length of work experience ($t(2070) = .79$, $p = .432$) and the length of service in the current place of work ($t(2035) = .64$, $p = .532$). However, in the PIP group the total work experience was

Figure 1. Proportion of subgroups in sample by group, sex, and level of education



on average 2.3 years longer than in the INT group, in which the average length amounted to 4.56 years ($t(1928) = 8.41$; $p < .001$, *95%CI* = 1.78–2.88). Similarly, the length of service in the current place of work was 3.6 years longer in the PIP group in which, on average, it amounted to 11.9 years ($t(2250) = 9.586$; $p < .001$, *95%CI* = 2.83–4.24).

Procedure

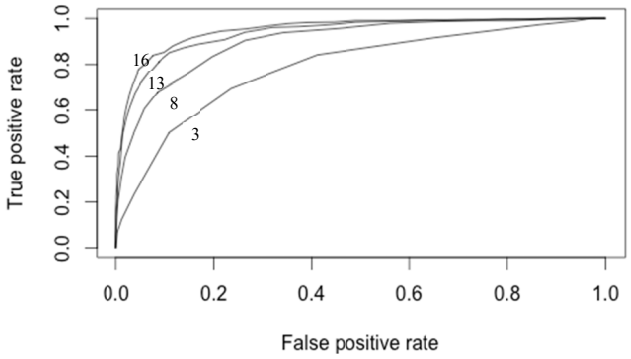
As an external criterion for the assessment of the quality of the examined tool versions, the study accepted the score calculated according to ICD-10 in the form of two values: encoded as 1 when there occurred minimum five scores at a level of at least four points (N = 1222) and as 0 when the number of high scores was not greater than two (N = 616). The remaining cases were excluded from further analyses (N = 620). The diagnosis based on the above criterion enabled, in the assessment of the quality of the scales with a short set of test items in relation to the original scale, using ROC curves (McNeil & Adelstein, 1976); (Patton, 1978). For a given version the diagnostic decision was confronted with the actual diagnostic result based on the criteria taken from ICD-10. Based on that, it was possible to determine sensitivity and specificity of particular versions of the tools and then determine the area under curve (AUC) which allowed assessing the quality of a given version of the test.

Sensitivity is defined as probability that a person afflicted with a disorder will be correctly identified in a test (*true positive rate*). In other words, this is the proportion of correctly identified individuals to the number of individuals with disorder. Specificity, on the other hand, is defined as probability that an individual without a disorder will be correctly identified in a test (*true negative rate*). In other words, it is the proportion of the number of correctly identified individuals to the number of people without disorder. The AUC score close to 1 would indicate a better

diagnostic power of a given tool – identifying individuals with the disorder as sick and individuals without the disorder as healthy and not making mistakes due the opposite classification.

The examples of curves for the LCWS scale where the score is based subsequently on 3, 8, 13 and 16 random items of the scale are presented in Fig. 2.

Figure 2. ROC curves for the score in the LCWS scale in the full version (16 items) and having 13, 8 and 3 items. The vertical axis indicates sensitivity and the horizontal axis (1-specificity)



In the study was used 4 methods for choosing the items that will be part of the questionnaires: 1) the coefficient of correlation with the general score of the scale (further referred to as *r*), 2) the factor analysis by the principal component method (further FA)[3], 3) logistic regression with the use of the criterion of belonging to the group of individuals who work excessively (further MRl) and 4) the item response theory (Hambleton, Swaminathan, & Rogers, 1991) with the use of the Graded Response Model (GRM – Samejima, 1969).

For the statistical technique based on the correlation coefficient, the result was calculated in the way which took into account the interfering effect of autocorrelation. Correlation of a given item with the general score of the scale was calculated for the adjusted general score from which the analysed test item had been excluded.

For the second statistical technique, in the factor analysis by the principal component method the study determined factor loadings for all test items of the questionnaire WEQ. Since all the items came from the full-form tool, the solution with 4 components was imposed in the factor solution. The results used the varimax rotation for the ease of interpretation and further analyses considered factor loadings only for the LCWS scale.

The method based on the logistic regression consisted in determining the values of the odds ratio (OR) for each item and choosing those which indicated the highest probability of the correct diagnosis.

In the case of the method based on IRT, for each item the value of the informative function was calculated using the parameters of difficulty and discrimination power determined on the basis of scores in GRM. The study assumed the difficulty range from -3 to 3 points.

The analyses comparing the statistical techniques also used two simple indicators describing the quality of particular items in the context of the order of their inclusion into the short version. The first one was the coefficient of order variability of the techniques: the square root of the sum of squared differences (SSD). Absolute differences of the order measured to what extent this sequence was different in the examined statistical techniques. The values close to zero indicated a high conformity of the techniques regarding the place on "the quality continuum" of a given item. On the other hand, the higher the values, the lower the conformity concerning the order. For example, item IT68P43, which on the basis of the logistic regression results should be first, and according to the remaining techniques was ranked 11th out of 16, gave SSD at a level of 17.3 points ($\sqrt{(1-11)^2+(1-11)^2+(1-11)^2}$). Another item, IT92P57, by two techniques ranked first and by two techniques second: SSD amounted to 2 points.

Also, the study determined the conformity measure for each of the examined techniques based on the differences between the sequence of particular items for a given technique and the averaged sequence of the 3 remaining techniques. The square root of the sum of squared differences for particular items divided by the number of test items enabled assessing the conformity of a given statistical technique with the remaining techniques within the entire questionnaire.

Instruments

The scale Lack of Control over Work of Working Excessively Questionnaire (WEQ) was chosen because of its high reliability (Cronbach's alpha was 0,89, SEM = 4.5), good discriminant power of items (from .43 to .67), and most importantly because we had objective (from ICD-10) information about health of participants.

Results

During the application of the above-mentioned four statistical techniques the sequence of particular test items regarding quality was ordered. Creating further short versions of a given scale took place by means of removing particular items according to the determined sequence. It was assumed that the shortest version of the scale would contain the best three test items (cf. Table 1).

So as to compare the order of items according to parameters from various statistical techniques, the sequence of particular items was treated as a recommendation from expert judges. For the examined LCWS scale of the WEQ questionnaire, the study obtained a statistically significant measure of the conformity of the sequences calculated with the use of Fleiss' kappa (Conger, 1980) and it amounted to .11. The result, however, is very low. Moreover,

[3] We agree with anonymous Reviewer that would be better to analyse questionnaires with Exploratory Factor Analysis than chosen Principal Component Analysis, but scores didn't differ much, and PCA is used more often, probably because of being the default setting in statistical packages like SPSS.

Table 1. The sequence of inclusion of items into the scale based on: 2) r – correlation with the general score, 3) FA – factor loadings obtained in the principal component analysis, 4) MRI – odds ratio in logistic regression, 5) IRT – informativeness level in the GRM model of the probabilistic approach. Absolute values of parameters of selected statistical techniques are put in brackets. Δ denotes difference in order on inclusion for the compared statistical techniques. SSD is measure of variability of inclusion of an item into the LCWS scale

LCWS	r	FA	MRI	IRT	Δ:r	Δ:FA	Δ:MRI	Δ:IRT	SSD
1	2	3	4	5	6	7	8	9	10
IT92P57	1 (0.60)	2 (.66)	2 (1,71)	1 (5,39)	.7	-.7	-.7	.7	2,00
IT126P78	3 (0.53)	3 (.62)	5 (1,38)	3 (3,61)	.7	.7	-2,0	.7	3,46
IT108P64	13 (0.40)	15 (.56)	14 (.92)	12 (2,44)	.7	-2,0	-.7	2,0	4,47
IT42P26	16 (0.34)	13 (.43)	13 (.90)	14 (1,83)	-2,7	1,3	1,3	.0	4,90
IT82P51	12 (0.49)	10 (.52)	8 (1,25)	10 (2,79)	-2,7	.0	2,7	.0	5,66
IT3P2	10 (0.43)	6 (.54)	7 (1,32)	8 (2,77)	-3,0	2,3	1,0	-.3	5,92
IT40P24	15 (0.43)	12 (.39)	12 (1,14)	15 (1,66)	-2,0	2,0	2,0	-2,0	6,00
IT123P76	8 (0.50)	5 (.56)	9 (1,17)	6 (3,18)	-1,3	2,7	-2,7	1,3	6,32
IT15P10	4 (0.52)	7 (.53)	3 (1,53)	7 (3,00)	1,7	-2,3	3,0	-2,3	7,14
IT86P54	7 (0.50)	4 (.58)	10 (1,17)	5 (3,16)	-.7	3,3	-4,7	2,0	9,17
IT113P68	14 (0.42)	14 (.53)	6 (.73)	13 (2,28)	-3,0	-3,0	7,7	-1,7	13,38
IT30P17	5 (0.50)	9 (.52)	16 (1,02)	9 (2,83)	6,3	1,0	-8,3	1,0	15,84
IT83P52	9 (0.44)	8 (.53)	4 (1,42)	16 (2,65)	.3	1,7	7,0	-9,0	17,29
IT68P43	11 (0.43)	11 (.44)	1 (2,07)	11 (2,56)	-3,3	-3,3	1.0	-3,3	17,32
IT31P18	6 (0.45)	16 (.57)	11 (.86)	4 (3,14)	4,3	-9,0	-2,3	7,0	18,63
IT22P13	2 (0.57)	1 (.66)	15 (1,04)	2 (4,57)	4,0	5,3	-13,3	4,0	23,15

Agreement		rho correlations			Average variability of the sequence for a given statistical method				
$\kappa = .11$	r	-.434	.203	.427					
z = 4,22	FA		-.294	.168	.71	.82	1,42	.84	1.04
p < .001	MRI			-.350					

SSD – variability of the sequence of inclusion into the scale expressed as the square root of the sum of squared differences, Δ: – difference in the rank of a given item between the rank in a given statistical technique and the average of the remaining techniques. Positive values indicate that a given statistical technique assigns to a given item a higher quality in comparison with the remaining statistical techniques. This value shows the size of the difference between the techniques in the assessment of the quality of an items.

the conducted conformity analysis of the sequence for statistical techniques did not provide any conclusive data; the highest conformity occurred between the correlation technique and the technique based on IRT (.427), and the lowest (in the sense of incompatibility): between the correlation technique and the factor technique (-.434).

It indicates that there is no universal quality of items according to which they could be ordered. Depending on the assumed statistical approach various items appear to be better from the perspective of the quality of the short form. This fact is confirmed by simple arithmetic conformity indicators: SSD and Δ, in which the lack of conformity between the statistical techniques regarding the sequence of particular items is noticeable.

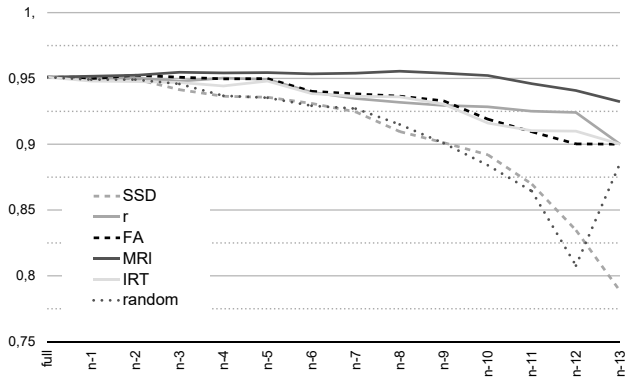
The obtained results point to the problem of indicating an optimum composition of the short version: each statistical technique highlight different questionnaire items as those which are more important. Using any of the statistical techniques, based on which the researcher takes decisions about excluding certain items from the tool, neither guarantees a success nor leads to a failure. The sequence in relation to the quality of particular items is assessed by different techniques in such a diverse way that, on the basis of it, it is impossible to draw conclusions and prepare recommendations.

In the second stage of the analyses the Authors focused on the assessment of the diagnostic quality of the short versions. To this end they used the AUC parameter of

ROC curves determined on the basis of a criterion which was external to the WEQ questionnaire, namely, the ICD-10 diagnosis.

Figure 3 presents the levels of AUC obtained for the LCWS scale. On the left there is a result illustrating the full scale and moving to the right one can see the result for the scale shortened by *k* items.

Figure 3. AUC for the LCWS scale shortened by subsequent questionnaire items



Legend: SSD – sequence according to the increasing root square of the sum of squared differences, r – sequence according to the decreasing coefficient of correlation with the general score, FA – sequence according to the decreasing values of factor loadings, MRI – sequence according to the decreasing values of the odds ratio, IRT – sequence according to the decreasing values of the informative function, random – a random sequence.

Along with shortening of the LCWS scale, its ability to correctly differentiate between the healthy and the sick decreases. This ability decreases most quickly when items are removed according to the SSD coefficient or randomly. The best stability of the diagnostic power of the scale can be obtained by the removal of items according to the sequence determined by logistic regression, which is a certain distortion resulting from the assumed method. Both AUC and logistic regression refer to the same criterion. The remaining methods of determining the sequence of removing items are placed between these extremes and none of them has a significant advantage over the others.

Discussion

The obtained results show a non-specific effect of the methods of questionnaire shortening. For example reliability of 8 items long scales was respectively .842, .837, .821, .836 for r, FA, MRI and IRT versions. For comparison for 10000 random versions of Cronbach's alpha computed on the same data mean reliability was $\alpha = .800$ with standard deviation SD = .01 (min = .758, max = .844, $\alpha > .83$ was 1.9%). This shows the advantage of statistical methods over the nonreflective shortening of the questionnaire scales. None of the examined statistical techniques proved universally better than the other techniques. Therefore, statistical techniques should be treated equally and selecting one of them comes down to the availability of analytical tools or a researcher's preference. Creating a short version on the basis of statistical parameters of items only is a task burdened with a considerable dose of uncertainty about the final result and with undefined stability of this solution.

Theoretical premises point to the advantage of less popular methods, i.e. analyses based on the IRT method or logistic regression. And if the former technique is in a way independent (it can be used having only a sufficiently big number of the observations of the results of a given tool), then logistic regression requires an external criterion for the assessment of the quality of particular items.

The presented method of assessing the diagnostic power of a tool based on the ROC curve method allows in a simple way comparing different versions of the tool and choosing the one which is marked by a lower "inclination" to make mistakes. Nevertheless, the Authors are from recommending any of the analysed techniques as sufficient. Creating a short version on the basis of the results collected with the use of the full version leads to obtaining a tool which has unknown diagnostic and psychometric properties. Without the content analysis and a complex psychometric analysis of the short version as a new test, what we receive is a research tool of undefined properties whose unreflective use will lead us up a diagnostic blind alley.

References

Bellak, L. (1955). *Short Form for TAT and CAT*. New York: CPS Inc.

Butcher, J.N., Kendall, P.C., Hoffman, N. (1980). MMPI short forms: CAUTION. *Journal of Consulting and Clinical Psychology*, *48*, 275–278.

Butcher, J.N., Hostetler, K. (1990). Abbreviating MMPI Item Administration. What Can Be Learned From the MMPI for the MMPI – 2? *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*(1), 12–21.

Canivez, G.L., Konold, T.R., Collins, J.M., Wilson G., (2009). Construct validity of the Wechsler Abbreviated Scale of Intelligence and Wide Range Intelligence Test: Convergent and structural validity. *School Psychology Quarterly 24*(4), 252.

Chushmir, L.H. (1985). Short-form scoring for McClelland's version of the TAT. *Perceptual and Motor Skills*, *61*, 1047–1052.

Conger, A.J. (1980). Integration and generalisation of Kappas for multiple raters. *Psychological Bulletin*, *88*, 322–328.

Costantino, G., Malgady, R.G., Vazquez, C. (1981). A comparison of the Murray TAT and a new thematic apperception test for Hispanic children. *Hispanic Journal of Behavioral Sciences*, *3*, 291–300.

Dahlstrom, W.G., Archer, R.P. (2000). A shortened version of the MMPI-2. *Assessment*, *7*, 131–141.

De Ayala, R. (1993). An Introduction to Polytomous Item Response Theory Models. *Measurement and Evaluation in Counseling and Development*, *25*(4), 172–189.

Doppelt, J.E. (1956). Estimated the full scale score on the Wechsler Adult Intelligence Scale from scores on four subtests. *Journal of Consulting Psychology*, *20*, 63–66.

Faschingbauer, T.R. (1974). A 166-item short form for the group MMPI: The FAM. *Journal of Consulting and Clinical Psychology*, *42*, 645–655.

Groth-Marnat, G. (2003). *Handbook of psychological assessment, Fourth edition.* New York: John Wiley & Sons.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, London, New Delhi: Sage Publications.

Jones, R.L. (1962). Analytically developed short forms of the WAIS. *Journal of Consulting Psychology*, *26*(3), 289.

Kaufman, A.S., Ishikuma, T., Kaufman-Packer, J.L. (1991). Amazingly short forms of the WAIS-R. *Journal of Psychoeducational Assessment*, *9*, 4–15.

Kincannon, J.C. (1968). Prediction of the standard MMPI scale scores from 71 items: The Mini-Mult. *Journal of Consulting and Clinical Psychology*, *32*, 319–325.

Levy, P. (1968). Short-form tests: A methodological review. *Psychological Bulletin*, *69*(6), 410–416.

McNeil, B.J., & Adelstein, S.J. (1976). Determining the value of diagnostic and screening tests. *Journal of Nuclear Medicine*, *17*(6), 439–338.

Overall, J.E., Gomez-Mont, F. (1974). The MMPI-168 for psychiatric screening. *Educational and Psychological Measurement*, *34*, 315–319.

Paluchowski, W.J., Hornowska, E., Haładziński, P., Kaczmarek, L. (2014). *Czy praca szkodzi? Wyniki badań nad kwestionariuszem nadmiernego obciążania się pracą.* Warszawa: Wydawnictwo Naukowe SCHOLAR.

Patton, D.D. (1978). Introduction to Clinical Decision Making. *Seminars in Nuclear Medicine*, *8*(4), 273–282.

Ryan, J.J., Ward, L.C. (1999). Validity, reliability, and standard errors of measurement for two seven-subtest short forms of the WAIS-III. *Psychological Assessment*, *11*(2), 207–211.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometrika Monographs*, *34*.

Satz, P., Mogel, S. (1962). An abbreviation of the WAIS for clinical use. *Journal of Clinical Psychology*, *18*, 77–79.

Silverstein, A.B. (1982a). Two- and four-subtest short forms of the Wechsler Adult Intelligence Scale-Revised. *Journal of Consulting and Clinical Psychology*, *50*, 415–418.

Silverstein, A.B. (1982b). Validity of Satz-Mogel-Yudin type short forms. *Journal of Consulting and Clinical Psychology*, *50*, 20–21.

Silverstein, A.B. (1990). Critique of a Doppelt-type short form of the WAIS-R. *Journal of Clinical Psychology*, *46*, 333–339.

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).

Stevens, M.R., Reilley, R.R. (1980). MMPI short forms: A literature review. *Journal of Personality Assessment*, *44*, 368–376.

The Psychological Corporation. (2001). *Wechsler Test of Adult Reading: Manual*. San Antonio, TX: The Psychological Corporation.

Ward, L.C., Selby, R.B., CIark, B.L. (1987). Subtest administration times and short-forms of the Wechsler Adult intelligence Scale-Revised. *Journal of Clinical Psychology*, *43*, 276–278.