STUDIES IN LOGIC, GRAMMAR AND RHETORIC 29 (42) 2012

How much credible are the responses obtained from an individual respondent in a non-repeated questionnaire survey: looking for practical methods with a statistical support

Izabela Chmiel[1], **Maciej Górkiewicz**[2]

[1] Department of Medical and Environmental Nursing, Faculty of Health Sciences, Jagiellonian University Medical College, Poland

[2] Department of Epidemiology and Population Research, Jagiellonian University Medical College, Poland

Abstract. All recognised psychometric methodologies, like the Classical Test Theory (CTT), the Item Response Theory (IRT), cognitive approaches, all in their essence are oriented on the population scale of the investigation. However, in daily practice, questionnaire inquiries are administered regularly only among a very limited group of people, frequently by very diverse professionals without any solid statistical background and usually with a clear practical purpose to support a decision on the course of therapy of an individual patient. Authors had some experience in both of the above domains. This paper was intended to remain on the borderline. First, it briefly discusses how the psychometricians put into practise the principle: proper instrument + proper procedure + proper attitude. The main focus was put on the demonstration of how to use several additional items in a questionnaire survey with an aim to verify the credibility of an individual respondent. The common pitfalls were illustrated with examples of analyses with use of easily available statistical procedures, like confidence intervals for proportion and the Friedman test for orderings.

Introduction

Historically, the psychometric and the statistical methodologies evolved in a close association, but without intense involvement with everyday psychological know-how. Many topics that are important for test users don't receive enough attention in psychometrics, so the question, what psychometrics can do for applied psychology, remains open [1–3]. Not a bit less vital can be the inverted question: what the questionnaire users can do for psychometrics [4]. However, the scope of this paper was limited to much simpler, practical question: how to enlarge our trust to the responses obtained from an individual respondent, basing above all on the common sense, supported with some relatively simple statistical procedures.

ISBN 978–83–7431–336–0 ISSN 0860-150X

Classical test theory (CTT) postulates an ideal situation, in which a perfect responder is always entirely eager to give the scores to each item of the reliable questionnaire of the simple linear structure, that can be expressed with an equation as (1) [5–6].

$$Y = \sum(X_i + e_i) = Y\hat{} + \sum(e_i); \quad i = 1, 2, \dots, K. \tag{1}$$

where: K – number of items in a single-scale questionnaire; $Y\hat{} = \sum(X_i)$ – actual result of the measurement, an estimate of the aggregate true score Y; X_i – responder's score to a particular i-th item; e_i – random error component, from definition of the expected value equal to 0.

Thus, in the frame of CTT, the prime attempt should be put on building a reliable questionnaires [7]. Nevertheless, the basic notions and indices of the CTT techniques, in these the factorial structure of the item set and Cronbach's alpha index of the internal consistency, became standard tools in every psychologist's tool kit [8].

Item Response Theory (IRT) generally didn't resign from the perfect respondent assumption, but it postulated more complex models then CTT. Unidimensional Rasch approach seemingly didn't modify the CTT formula (1), so an aggregate score is estimated in the same way, as the sum of the scores obtained by the items in the questionnaire. The difference here has a structural nature for it was assumed that true scores of items are put in linear order [9]. It is easy to notice that with respect to ordered chance to be censored, the error components of the particular items have their expected values generally not equal to 0. Nevertheless, if the postulated ordering of the true scores exist in the real world, they can be easy estimated, because of the relatively small probabilities of the contradictory actual scores [10]. In a doubtful case it seems to be more appropriate to resign from the Rasch modeling, than to adjust the essentials of a model to the actual data [11]. The Structural Equations Models (SEM) technique can connect, in various ways, several equations like (1). Moreover, SEM created an opportunity to include into a joint model many variables, latent as well as manifest, like features of the responders, and attributes of a survey [12–14].

The cognitive approaches, contrary to CTT and IRT, gave attention to the psychological aspects of the questionnaire inquires, before all in terms of responder's ability and readiness to provide the honest answers to the questionnaire items [15–16]. Practical recommendations how to organize a questionnaire survey are based here on real-world observations [17], with correspondence to known psychological approaches, such as the false memory phenomenon [18], the cognitive-affective models of goal-setting [19], the model of planned behaviour [20], or the attitude-social influence-efficacy (ASE) paradigm [21].

The professional developers of the questionnaires addressed to broad target must create a balanced amalgam of some practical modus operandi. For instance, a potential user of the known SF-36 questionnaire, besides detailed handbook how to carry out a survey and all subsequent calculations, can get a lot of authorized and very useful information [22], on the factorial structure of SF-36 [23], on Rasch models for SF-36 [24–25], on path models for SF-36 [26]. In addition, the numerous disinterested scientific reports are available, among other with regard to nonstandard subjects [27], or to nonstandard procedure [28]. The adaptation of a standard questionnaire to other (nonstandard) population needs a special prudence [29]. However, this seems to be less risky for the researcher than trying to create ad hoc ones own new questionnaire, especially without proper psychological background [30].

The rest of this paper was organized as follows. Brief examples on standard adaptation and validation procedures from the authors own studies are given in the chapter: The ground rule: proper instrument + proper procedure + proper attitude. In the next chapter: Statistical supports for credibility of an individual respondent, the method based on comparisons between responder's opinion versus a corresponding pattern is proposed. The two exemplary patterns, that is the typical relationship between the inclination to guess Yes-or-No answer versus the level on the disagreement in the matter [31], and the standard ordering of the items of the physical functioning (PF) scale in SF-36 questionnaire were obtained from our previous studies of Polish groups [32].

The ground rule: proper instrument + proper procedure + proper attitude

The developers of the wildly used questionnaires have made a vital attempt to be in a reasonable agreement with all recognized approaches to psychometrics. With respect to the applied surveys the common recommendations for the researchers, how to enlarge their chance to obtain valid results, can be summarized as follows [33]:

1. draw a representative sample of responders from a population under study;
2. use a standard questionnaire;

3. apply a standard procedure if possible;
4. confirm a sufficient similarity between standard population and a population under study.

In spite of a broad use, the notion of a standard questionnaire hasn't any formal definition. The usual obligatory demands for a standard questionnaire included: available detailed scientific report from a large scale confirmatory survey, authorized handbook with instructions how to carry out justifiable surveys with use of this questionnaire, and desirably, at least several research reports from surveys made with use of this questionnaire in various circumstances and by different research teams. It should be emphasized once again, that the practicality and efficacy of any standard questionnaire and of the standard procedure of its use, both were proved jointly by their developer with respect to some accurate specified standard population. Thus, if an actual survey was made at some other population, then the sufficient similarity between standard population and the population under study must be proved thoroughly [34]. In case of need, it should be proved that the all identified differences were irrelevant in the matter, for instance for the fertility behaviour [35]. However, a great number of the potentially influential variables makes very likely the occurrence of the Simpson paradox [36].

The authorized handbook [22] provided the detailed recommendations how to carry out a survey with the standard SF-36 questionnaire, and then, how to make calculations and interpret the results. All 36 items of the SF-36 questionnaire produce only 9 variables (health-related quality of life domains): GH – general health; HT – change in health; VT – energy/vitality; MH – mental health; RP – role limitation-physical; SF – social functioning; BP – bodily pain; RE – role limitation-emotional, and PF – physical functioning. The raw SF-36 data should be standardized with a range of 0–100% separately for each the above 9 scales. For the purposes of the confirmatory analyses the two standard populations where characterised with their estimates of the mean values and standard deviations for the three domains: PF = 83.29±23; RP = 82.51±25; VT = 58.31±20 for the USA general population, and PF = 83.9±11.6; RP = 72.4±5.1; VT = 64.5±5.7 for the Finnish general population.

The authorized recommendations [22] were respected rigorously at the thesis stage [33], however some additional analyses were applied. The fundamental presumption that the study group can be considered as representative, at least for Polish convalescents after successful clinical therapy against *acute pancreatitis,* was supported by several arguments. The initial sample included all of the 422 patients hospitalised for *acute pancreatitis* at the 1st Department of General Surgery at the Jagiellonian University of Krakow (Poland) from 2000 to 2006. The only four exclusion criteria were used: age: < 18 years or > 70 years (66 excluded); death (34 excluded); non complete clinical data (20 excluded), complication with other illness (36 excluded). The standard procedure for the mail survey was applied with proper thoroughness. The standard Polish version of SF-36 questionnaire with standard instructions was mailed to all of the 266 non-excluded survivors. A covering letter accompanying each questionnaire included also the explanation of the survey purpose and of the possible health benefits for the respondent. A phone consultation in completing the form, if needed, was offered. Nevertheless, the $N = 124$ participants didn't return an answer, but $N = 142$ survivors (81 men and 61 women) returned acceptably completed forms. The three clinical types of disease were represented at the study sample with appropriate proportion: 61:41:40. The response rate RR = 142/266 = 53.4% was acknowledged as sufficient for the mail survey. Moreover, the clinical and demographic data for non-responders and responders were quite similar, so the adjusting for non-response was unnecessary. The assumptions of normality of the scores for the 9 particular health-related quality of life domains, as measured with the SF-36 questionnaire at the study group, were supported with moderate values of skew and ranged from skew $= -0.52$ to skew $= 0.24$, and of the kurtosis varied between kurtosis $= -1.12$ and kurtosis $= 0.43$. Consequently, the confirmatory analyses of the data reliability and validity were executed predominantly at the frame of the classical test theory (CTT) with the use of the parametric procedures. The proper correlation structure of the raw SF-36 data was confirmed for each domain separately not only under the criterion that Cronbach's alpha > 0.7 but also under the criterion that each item is in a quite strong correlation with the summary score of its domain, but is in relatively weaker correlation with any other item. The concurrent validity was confirmed on the base of estimates for the study group: PF $= 64.5\pm27.1$; RP $= 59{,}0\pm30{,}9$; VT $= 52.5\pm16.8$; and as well, on the base of estimates for the Finnish convalescents also after *acute pancreatitis*: PF $= 83.0\pm21.6$; RP $= 69.4\pm27.8$; VT $= 60.4\pm23.4$ [37]. It was easy to notice, that the study group didn't differ significantly with respect to PF, RP and VT domains from any of the above groups, because all considered differences between mean values were less than their standard deviations of the study group. Beyond the obligatory recommendations, the study data were reanalysed at several parallel studies with the use of somewhat more advanced methodology [22]. The postulated linear ordering of the items of PF scale in the study sample was confirmed with the Rasch methodology, and then used in comparative analyses with the aim to confirm the concurrent validity of the study data [32]. In the study data the significant regression was detected between mean scores of the SF-36 domains and their standard deviations, SDˆ $= -0.043 + 0.524$*mean; $R = 0.705$ with statistics $F = 6.9$; $p = 0.03$. In such a situation, the multiple comparisons procedures [38], and the bootstrap [39], were used with the aim to get an additional support to previous conclusions in the matter, based on parametric procedures. The suitability of the applied clinical classification of the patients menaced with *acute pancreatitis* was confirmed also in terms of the propensity score [40]. The informative links between age and gender on one side, and the chosen SF-36 domains on the other side were estimated [41–42].

The analyses (cited in this chapter) provided strong support to conclude that the study group is representative, at least for Polish convalescents after *acute pancreatitis*, that in general the members of the study group gave trustworthy scores to items of the SF-36 questionnaire. These findings raise the possibility that the data obtained in this group with other, nonstandard questionnaires, can be considered as a source of valid information. Basing on this conviction, the health behaviours of convalescents after *acute pancreatitis* were classified [43], recommendations on needed psychoeducational intervention for convalescents were proposed [44–45], and the proposal to school's health education were suggested [46].

Quite analogous approach, following scrupulously the recommendations of the developers of the standard questionnaire, but not neglecting the parallel analyses with other methodologies, were applied in our studies on adopting the known CES-D questionnaire [47], and in adopting the physicians' career satisfaction questionnaire [48]. The forward-backward translation procedure was applied with special attention to a dogma, that the more the respondents are emotionally invested in the item, the more likely those emotions will influence their scores. Concurrent validity of CES-D was proved by comparison with scores obtained through the well-known Beck Depression Inventory. The responders from the sample included 3544 permanent residents of Krakow (Poland), recruited from the HAPIEE Study (Health, Alcohol and Psychosocial factors In Eastern Europe, url=http://www.ucl.ac.uk/easteurope/hapiee.html). Besides, the two above studies the concurrent cross-cultural validity was confirmed by the similarity between standard factorial structure and the one estimated for the Polish version of the adopted questionnaire.

The endeavour to introduce a novel questionnaire creates new fundamental challenges for developers. In such case, the structural equation modelling (SEM) in the frame of item response theory (IRT) should be preferred. This allows exploration in compound of the truly multivariate models, where multiple independent variables can influence multiple intermediate variables in the prediction of the final effects of the antecedent variables. In the study [49], two simple unidimensional scales, as described by the equation (1), were examined. First, the questionnaire uses only three items to measure the latent variable named motivation-to-work, and the other one uses eleven items to measure the latent variable named attitude-to-patients. The SEM model linked one latent variable to the other. In result, the unidimensionality of both questionnaires under study, and the significant correlation between the considered latent variables, $R^2 = 0.78$; $p < 0.001$ was confirmed simultaneously. In continuation [50] the use of some easy available data of a candidate nurse as a substitute to the questionnaire review with above questionnaires was considered. However, it should be emphasized that this technique seems to be useful for managers looking for suggestions how to develop patient-friendly staff in a rather long perspective, but not to evaluate an individual candidate. In the study [51] on the false memory, the latent variable named Model, expressed an inclination of the respondent to invent the “ad hoc” explanations in spite of insufficient information. It was proved that a simple linear model like (1) should be rejected from consideration. In the final quasi-linear SEM model the six paths, each significant, at least on the level of $p < 0.034$, connected four independent variables (gender, ratio of the true answers, and ratios of two kinds of the wrong answers) and the latent variable Model into a complex relationship, and visibly different from a simple model (1).

Statistical supports for credibility of an individual respondent

The question as such, had been a vital issue in almost all real-life domains. However, in medicine the problem, of how far patient’s opinions may be trusted, has its special significance. Generally, there is an agreement between regulatory authorities and the research community that patient-reported outcome (PRO) assessment in health care should proceed from a strong conceptual basis, with rationales clearly articulated in advance concerning what is to be measured and how this is to be accomplished, with greater awareness to recall bias and degrees of psychometric validation [52–55]. It should be recognised that patients’ and their care-providers’ views can show some discrepancy, especially with regard to the course of rehabilitation and other long-time care, therefore, the interviews carried out by other persons from outside seem to be indispensable here [46, 56–57]. Before choosing a validation method the two crucial prerequisites should be considered thoroughly:

(i) the hypothetical source of false answers: unplanned random answering versus intentional (maybe: to some extent unconscious) play-acting or pretending;

(ii) the anticipated meaning of a patient's opinion: expert's report versus subjective conviction or impression.

As to the first of the above issues (i), the strategy of random answering can be easy modelled with the use of some commonly applied standard distribution. The strategy of inventing fictitious self-image is generally more difficult to unveil, especially without any clear concept of a possible pattern or a scale of a self-worth underlying this strategy. In this study we attempted to disregard this problem by using nonparametric statistical procedures.

As to the second issue (ii), expert's opinion can be verified directly by comparison with real world occurrences, and with opinions of other experts. This problem has great practical relevance, and plentiful literature on the matter, nevertheless, it wasn't included in the scope of this paper. The dishonesty of subjective conviction is generally more difficult to reveal. In this paper we suggest the use of a characteristic pattern, that is a typical relationship between variables measured in a questionnaire survey at some postulated populations. The proposed validation techniques were aimed to recognise a responder either as outsider or as a member of these populations. The practical difficulties with the use of the two patterns, that is the typical relationship between the inclination to guess Yes-or-No answer versus level on the disagreement in the matter [31], and the standard ordering of the items of the physical functionning (PF) scale in SF-36 questionnaire [32], were explained in this paper with the exemplary statistical calculations.

The method for evaluating an individual inclination to guess Yes-or-No answers [31], was originally developed to examine the members of a small group of experts. In the experiment the role of anonymous experts played $N = 84$ graduated nurses. The questionnaire included mixture of controversial items with different levels of disagreement in literature. The two kinds of items with dichotomous decisive Yes-or-No answer were used: the $K = 31$ items allowed apparently only decisive answer, but $L = 44$ items permitted explicitly also the third I-don't-know option of a answer. In such a way the questionnaire created series of two seemingly equivalent decision situations. In the first situation the participants, aimed to avoid a decisive answer, giving neither Yes nor No answer. But in the second situation they can choose freely an additional option I-don't-know. It was proved that, even in the anonymous survey, the same participants avoided the decisive Yes or No answer, significantly less often in the first situation, only 3 times at $N \cdot K = 84 \cdot 31 = 2604$ answers, what leads to a proportion $Pr_1 = 3/2604 = 0.0012$; 95%CI: $Pr_1 < 0.003$; than in the second situation, 608 times at $N \cdot L = 84 \cdot 44 = 3696$ answers, what leads to much greater proportion $Pr_2 = 608/3696 = 0.165$; and confidence interval 95%CI: $0.153 < Pr_2 < 0.177$. The odds ratio $OR = Pr_2/Pr_1 = 142.8$ with a confidence interval $74.0 < OR < 275.7$. Moreover, the strong log-linear relationship (2), $p < 0.01$; between the proportion of Yes versus No answers and the frequency of the I-don't-know answers at the $L = 44$ items with this option was observed. It is easy to notice in equation (2), that odds ratio $OR_{\text{I-don't-know/I-know}}$ obtained its maximal value for $\text{Ln}|OR_{\text{Yes/No}}| = 0$; that is in situation when (frequency of Yes) = (frequency of No).

$$\text{Ln}(OR_{\text{I-don't-know/I-know}}) = 0.526 - 0.943 \cdot \text{Ln}|OR_{\text{Yes/No}}| \tag{2}$$

where:

$OR_{\text{I-don't-know/I-know}} =$
$=$ (frequency of I-don't-know)/(frequency of either Yes or No);

$OR_{Yes/No} =$ (frequency of Yes)/(frequency of No).

It seems that the design of a verifying experiment should be limited here to three binary variables only, that is: level of agreement in a standard population with regard to choice between Yes versus No answer (Agreement = low vs. Agreement = high), encouragement to I-don't-know answer (Option I-don't-know = offered vs. Option = hidden), answer = decisive versus answer = ambiguous. In result the set of $N = 60$ respondent answers to the verifying items can be summarised as a 3D table of frequencies, like [Tab. 1]. The null hypothesis that the respondent provided his answers independently from item's values of variables Agreement and Option can be easily proved with calculator for Fisher exact test, available on-line [58]. It should be noted that in spite of relatively large number $N = 60$ of the verifying items, the estimated significance of the null hypothesis was quite near to $p = 0.05$. Thus, the use of the discussed method for evaluating an individual inclination to guess Yes-or-No answers, seems to be useful only in a situation if a researcher is truly interested in the viewpoint of a respondent on almost all of the verifying items.

Tab. 1. Exemplary data on an individual inclination to guess Yes-or-No answers

Option I-don't-know	Yes:No Agreement = low	Yes:No Agreement = high	total
Offered	$N_{\text{ambiguous}}/N_{\text{decisive}} = 3/7$	$N_{\text{ambiguous}}/N_{\text{decisive}} = 3/12$	6/19
Hidden	$N_{\text{ambiguous}}/N_{\text{decisive}} = 1/14$	$N_{\text{ambiguous}}/N_{\text{decisive}} = 0/20$	1/34
Total	$N_{\text{ambiguous}}/N_{\text{decisive}} = 4/21$	$N_{\text{ambiguous}}/N_{\text{decisive}} = 3/32$	7/53

N_{decisive} – number of either Yes or No answers;
$N_{\text{ambiguous}}$ – number of other answers;
Yes:No Agreement = low
    if in standard population Probability(Yes) ≈ Probability(No);
Yes:No Agreement = high
    if in standard population |Probability(Yes) – Probability(No)| > $\frac{1}{2}$;
Null hypothesis H_0:
    the same probabilities in each cell $Pr(N_{\text{ambiguous}}/N) = 7/(7+53) = 7/60$;
Fisher exact test:
    two-sided mid-significance $p = (0.037 + 0.033)/2 = 0.035$; reject H_0;
Pearson χ^2 test don't valid:
    $\chi^2 = 7.28$; *df* = 3; significance $p = 0.065$; don't reject H_0.
On-line calculator: http://www.quantitativeskills.com/sisa/statistics/fiveby2.htm

The other, a strongly ordered pattern of J objects, $O_1 < O_2 < \ldots < O_J$, for recognising a responder either as an outsider or as a member of some assumed population can be easily constructed basing on established standard ordering, for instance on ordering of ten items of the physical functioning in the SF-36 questionnaire [32]. Several known procedures can be used to prove the level of concordance between estimated responder's ordering versus ordering assumed in a pattern [59]. However in this paper, for significant reasons it was suggested to apply the procedure of pair-wise comparisons with some other fixed object O_x from inside a pattern, with further use of the Friedman test [60]. The first reason is that, under analogous exertion for a respondent, the comparisons usually lead to more reliable estimates than rankings [61]. Moreover, the needed sample size for the Friedman test begins here from $J = 7$ objects in an ordered pattern and only two or three other objects O_x [60]. Thus, the proposed way of verification of a respondent can be made with no more than 21 verifying items, added with this purpose to the core questionnaire. The logic and computational details of the Friedman test are described in [60]. All computations are straightforward, the formulas (3–5) can be easily implemented in any universal spreadsheet, in case of necessity with the use of only basic arithmetical operations. Moreover, the host [60] submits on-line access to the user-friendly calculator for $K = 3$ and $K = 4$ initial rankings, which performs automatically all further calculations of the Friedman test.

In [Tab. 2] for each object from a given pattern in the section named raw initial ranks there were shown results of the three separate evaluations, named X, Y, Z, obtained with the 9-level Likert scale, from score = 1, by step = 1, up to score = 9. For instance, object O_1 obtained scores $x_1 = 1$, $y_1 = 2$ and $z_1 = 3$; object O_2 obtained scores $x_2 = 2$, $y_2 = 3$ and $z_2 = 4$; and so on, up to object O_7 with its scores $x_7 = 7$, $y_7 = 8$ and $z_7 = 9$.

With the aim to verify hypothesis H_0', that the evaluations X, Y and Z didn't differ with respect to this pattern, at the beginning separately for each object its raw evaluations were changed with their relative ranks. For instance, the ranks of the object O_1 were transformed into relative ranks $x_1' = 1$, $y_1' = 2$ and $z_1' = 3$; because its raw ranks are ordered: $x_1 < y_1 << z_1$. Analogously, for each other object, its minimal raw evaluation was transformed into relative rank = 1; its intermediate raw evaluation into relative rank = 2; and its maximal raw evaluation into relative rank = 3. The Kendall's coefficient of concordance W was estimated with formula (3) as $W = 1$. The test statistic Q was estimated with formula (4) as $Q = 14$. Because the data sample was sufficiently large, the distribution of the test statistic Q can be considered as a close approximation of the chi-square distribution with degree of freedom equal to $df = K - 1$ [60]. Therefore, the significance of the null hypothesis H_0 was estimated with formula (5) as $p = 0.0009$, manifestly smaller than $p = 0.05$. Thus, the null hypothesis H_0' should be rejected without any serious doubt. It should be concluded that the evaluations X, Y and Z did differ significantly with respect to a pattern under investigation.

$$W=12\cdot\textstyle\sum_k \left(\sum \text{relative.rank}_j|k\right)^2/J^2\cdot K\cdot(K^2-1))-3\cdot(K+1)/(K-1), \quad (3)$$

$$Q = J \cdot (K-1) \cdot W, \quad (4)$$

$$p(Q) = p(\chi^2 = Q)|(df = K-1) \quad (5)$$

where: $j = 1, 2, \ldots, J$; J – number of objects under evaluation; $k = 1, 2, \ldots, K$; K = number of ways of evaluation; relative.rank$_j|k$ – relative rank of j-th object under k-th way of evaluation.

With the aim to verify somewhat different hypothesis H_0'', that the evaluations X, Y and Z were generated by the same latent ordering of the compared objects, at least with an insignificant random error, at the beginning each raw initial evaluation X, Y and Z separately should be transformed into standardized ranks. For instance, object O_1 got standardized rank $y''_1 = 1$ because its raw score $y_2 = 2$ was a minimal Y score, but object O_7 got standardized rank $y''_7 = 7$ because its raw score $y_2 = 8$ was a maximal Y score among the all $J = 7$ of Y scores under investigation. The standardized ranks of the remaining objects and the remaining ways of scoring were defined as usual. Subsequently, the standardized ranks were processed in the same manner as the raw evaluations formerly. For instance, the standardized ranks of the object O_1 were transformed into relative ranks $x'_1 = 2$, $y'_1 = 2$ and $z'_1 = 2$; because its standardized ranks are just the same: $x_1 = y_1 = z_1 = 2$. Thus, the Kendall's coefficient of concordance W was estimated with formula (3) as $W = 0$. The test statistic Q was estimated with formula (4) as $Q = 0$. The significance of the null hypothesis H_0 was estimated with formula (5) as $p \approx 1.0$, manifestly greater than $p = 0.05$. The null hypothesis H''_0 should be accepted without any serious doubt. It should be concluded that the scores X, Y and Z were generated by the same latent ordering of the compared objects.

Tab. 2. Friedman test for exemplary data of K = 3 orderings without ties

pattern	raw initial ranks			relative row ranks			standardized ranks			relative row ranks		
object	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
1	1	2	3	1	2	3	1	1	1	2	2	2
2	2	3	4	1	2	3	2	2	2	2	2	2
3	3	4	5	1	2	3	3	3	3	2	2	2
4	4	5	6	1	2	3	4	4	4	2	2	2
5	5	6	7	1	2	3	5	5	5	2	2	2
6	6	7	8	1	2	3	6	6	6	2	2	2
7	7	8	9	1	2	3	7	7	7	2	2	2
sum	–	–	–	7	14	21	–	–	–	14	14	14

$B = J_2 \cdot K \cdot (K^2 - 1)) = 7 \cdot 7 \cdot 3 \cdot (3 \cdot 3 - 1) = 1176;$
$C = 3 \cdot (K + 1)/(K - 1) = 3 \cdot (3 + 1)/(3 - 1) = 6;$
for raw initial ranks:
$A = 12 \cdot \sum(\text{sum}^2) = 12 \cdot (7 \cdot 7 + 14 \cdot 14 + 21 \cdot 21) = 8236$; $W = A/B - C = 1$;
$Q = J \cdot (K - 1) \cdot W = 7 \cdot (3 - 1) \cdot 1 = 14$; $df = K - 1 = 2$; $p(\chi^2) = 0.0009$;
conclusion: raw initial ranks of the $J = 7$ objects from a pattern differ significantly;
for standardized initial ranks:
$A = 12 \cdot \sum(\text{sum}^2) = 12 \cdot (14 \cdot 14 + 14 \cdot 14 + 14 \cdot 14) = 7056$; $W = A/B - C = 0$;
$Q = J \cdot (K - 1) \cdot W = 0$; $df = K - 1 = 2$; $p(\chi^2) \approx 1.0$;
conclusion: standardized initial ranks of the $J = 7$ objects from a pattern don't differ significantly.

It should be emphasized that the manifestly opposite conclusions for the above null hypotheses H'_0 and H''_0 were both obtained in the approved manner with the same Friedman test on the base of the exactly the same raw data. This occurrence exemplified the first devious trap that was covered in the Friedman test methodology: a researcher should distinguish the real meaning of comparing the raw evaluations versus comparing the standardized ranks of these evaluations.

In the proposed procedure all elements of a pattern, $O_1 < O2 < \ldots < O_J$, $J = 7$, are presented separately, in random sequence, and a responder is asked to compare a presented object with also separately presenting the three other fixed objects O_i; $i = X, Y, Z$; from inside a pattern, using 5-level Likert scale, score 1: $O_i << O_j$; score 2: $O_i < O_j$; score 3: $O_i \approx O_j$; score 4: $O_i > O_j$; score 5: $O_i >> O_j$; were: relation $<<$ denotes a judgement "definitely less ..."; $<$ denotes "rather less ..."; $\approx$ denotes "rather not different ...". Because number J of objects is greater than the number of a Likert scale levels, the occurrence of ties (the same scores for some objects) is inevitable here.

The real-life exemplary data were shown and analysed in [Tab. 3]. As above, in [Tab. 2], also the two different null hypotheses were verified in [Tab. 3]:

H_0': The ranks of the objects assumed in the pattern and all three raw initial ranks didn't differ significantly;

H_0'': The ranks of the objects assumed in the pattern and all three standardized ranks didn't differ significantly.

The significance of the null hypothesis H_0 was estimated here with the formula (5) as $p = 0.011$, manifestly smaller than $p = 0.05$. For that reason, the null hypothesis H_0' should be rejected without any serious doubt. It should be concluded that the raw initial ranks cannot be generated by the same latent ordering of the compared objects as is assumed in the pattern. The significance of the null hypothesis H_0'' was estimated with formula (5) as p 0.99, manifestly greater than $p = 0.05$. For that reason, the null hypothesis H_0'' should be acknowledged without any serious doubt. It should be concluded that all three standardized ranks can be generated by the same latent ordering of the compared objects as is assumed in the pattern.

It should be emphasized that the manifestly opposite conclusions for the above null hypotheses H_0' and H_0'' both were obtained in the approved manner with the same Friedman test on the base of exactly the same raw data. This occurrence exemplified the second devious trap that was covered in the Friedman test methodology: a researcher should distinguish the real meaning of the comparing the raw evaluations defined with the various Likert scales versus comparing the standardized ranks of these evaluations defined with exactly the same Likert scales (that is the same origin and the same number of levels at all used Likert scales).

Tab. 3. Friedman test for exemplary data of K = 4 orderings with some ties

pattern	raw initial ranks			relative row ranks				standard ranks			relative row ranks			
object	X	Y	Z	patt	X	Y	Z	X	Y	Z	patt	X	Y	Z
1	2	1	1	2	4	2	2	1.5	1.5	1.5	1	3	3	3
2	4	2	2	2	4	2	2	3.5	3.5	3.5	1	3	3	3
3	2	1	1	4	3	1.5	1.5	1.5	1.5	1.5	4	2	2	2
4	5	4	3	2.5	4	2.5	1	6	5.5	5.5	1	4	2.5	2.5
5	4	2	2	4	3	1.5	1.5	3.5	3.5	3.5	4	2	2	2
6	5	4	3	4	3	2	1	6	5.5	5.5	3.5	3.5	1.5	1.5
7	5	5	4	4	2.5	2.5	1	6	7	7	3	1	3	3
sum	–	–	–	22.5	23.5	14	10	–	–	–	17.5	18.5	17	17

$B = J^2 \cdot K \cdot (K^2 - 1)) = 7 \cdot 7 \cdot 4 \cdot (4 \cdot 4 - 1) = 2940$;
$C = 3 \cdot (K + 1)/(K - 1) = 3 \cdot (4 + 1)/(4 - 1) = 5$;
for raw initial ranks:
$A = 12 \cdot \sum(\text{sum}^2) = 12 \cdot 1354.5 = 16254$; $W = A/B - C = 0.529$;
$Q = J \cdot (K - 1) \cdot W = 7 \cdot (4 - 1) \cdot 1 = 11.1$; $df = K - 1 = 3$; $p(\chi^2) = 0.011$;
conclusion: raw initial ranks of the $J = 7$ objects from a pattern differ significantly;
for standardized initial ranks:
$A = 12 \cdot \sum(\text{sum}^2) = 12 \cdot 1226.5 = 14718$; $W = A/B - C = 0.006$;
$Q = J \cdot (K - 1) \cdot W = 0.13$; $df = K - 1 = 3$; $p(\chi^2) \approx 0.99$;
thus, standardized initial ranks of the $J = 7$ objects from a pattern don't differ significantly.

Discussion and conclusions

The patient is the primary recipient of treatment, so it is an urgent need to recognize the patient's own perspective on the illness experience and the effects of therapy, as necessary and unique complement to all professional's evaluations. Therefore, in a daily medical practice, questionnaire inquiries are administered regularly with clear practical purpose to support a decision on the course of therapy in very limited groups of patients. This study was focused on how to make results of the questionnaire examinations more reliable and easily understandable to health workers and other professionals with a limited background in the psychometric and statistical methodology.

Generally, this paper proposed an intuitive, yet statistically precise approach to applied questionnaire examinations. The first topic, 'The ground rule: proper instrument + proper procedure + proper attitude', corresponded to typical simple way of reasoning: we can trust in the data obtained from an individual respondent, because these data are only a fragment of the whole data set from questionnaire survey of the proved reliability and validity. This approach can fail, particularly in situation if an individual respondent under examination gave the false answers, but the final scores of these answers satisfiedthe usual formal criterions. Therefore, in the frame of the second topic, ‘Statistical supports for credibility of an individual respondent’, a fresh and innovative approach to task of recognising an individual respondent either as a typical member or as an unusual member of the homogenous group is suggested. The proposed methodology corresponded to somewhat more sophisticated way of reasoning: we can trust in all data obtained from an individual respondent, because the answers of this respondent to a set of the verifying items are in a general agreement (or: in a close agreement) with the acknowledged pattern. The use of known Friedman test is then recommended. The Friedman test can be considered as a nonparametric two-way analysis on ranks. In spite of all its advantages, in practice the Friedman test was not often used, maybe because of the two devious traps that lurk there for an inexperienced researcher. For this reason, the Friedman test procedure, and the associated common misunderstandings were thoroughly explained in this paper with the exemplary data showed in [Tab. 2] and [Tab. 3]. The two exemplary patterns, that is the typical relationship between the inclination to guess Yes-or-No answer versus level on the disagreement in the matter, and the standard Rasch ordering of the items in an applied questionnaire, were based on Authors’ own previous studies in Polish groups.

REFERENCES

[1] Sijtsma K., Future of Psychometrics: Ask What Psychometrics Can Do for Psychology, Psychometrika, 77 (1), pp. 4–12, 2012.

[2] Sijtsma K., Reliability Beyond Theory and Into Practice, Psychometrika, 74 (1), pp. 169–173, 2009.

[3] Borsboom D., The Attack of the Psychometricians, Psychometrika, 71 (3), pp. 425–440, 2006.

[4] Gimeno-Santos E., Frei A., Dobbels F., et al., Validity of instruments to measure physical activity may be questionable due to a lack of conceptual frameworks: a systematic review, Health and Quality of Life Outcomes, 9 (86), 2011. http://www.hqlo.com/content/9/1/86

[5] StatSoft, Inc. Reliability and Item Analysis, in: StatSoft, Inc. (2012). Electronic Statistics Textbook, Tulsa, OK: StatSoft,
WEB: http://www.statsoft.com/textbook/.2012

[6] de Klerk G., Classical test theory (CTT), In Born M., Foxcroft C. D. & Butter R. (Eds.), Online Readings in Testing and Assessment, International Test Commission, 2008. http://www.intestcom.org/Publications/ORTA.php

[7] Streiner D. L., Norman G. R., Health measurement scales a practical guide to their development and use, Oxford University Press, Inc., New York, 1989.

[8] Sijtsma K., On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha, Psychometrika, 74 (1), pp. 107–120, 2009.

[9] Fischer G. H., Molenaar I. W., Rasch Models – Foundations, Recent Developments, and Applications, Springer-Verlag, Berlin, 1995.

[10] Masters G. N., A Rasch model for partial credit scoring, Psychometrika, 47, pp. 149–173, 1982.

[11] Tennant A., Penta M., Tesio L., et al., Disordered Thresholds: An Example from the Functional Independence Measure, Rash Measurement Transactions, 17 (4), pp. 945–948, 2004. http://www.rash.org/rmt/rmt174a.htm,

[12] Chang H. H., Wang C., Book Review [M.D. Reckase (2009) Multidimensional Item Response Theory. New York: Springer], Psychometrika, 76 (3), pp. 504–506, 2011.

[13] Willse J. T., Goodman J. T., Comparison of Multiple-Indicatorrs, Multiple-Causes – and Item Response Theory-Based analyses of Subgroup Differences, Educational & Psychological Measurement, 68 (4), pp. 587–602, 2008.

[14] StatSoft, Inc.Structural Equation Modeling., in: StatSoft, Inc. (2012). Electronic Statistics Textbook, Tulsa, OK: StatSoft.
WEB: http://www.statsoft.com/textbook/

[15] Collins D., Pretesting survey instruments: An overview of cognitive methods, Quality of Life Research, 12, pp. 229–238, 2003.

[16] Mislevy R. J., Verhelst N., Modeling item responses when different subjects employ different solution strategies, Psychometrika, 55, pp. 195–215, 1990.

[17] Rimm E. B., Stampfer M. J., Colditz G. A., et al., Effectiveness of various mailing strategies among nonrespondents in a prospective cohort study, Am J Epidemiol, 131, pp. 1068–1071, 1990.

[18] Gerrie M. P., Belcher L. E., Garry M., Mind the gap: false memories for missing aspects of an event, Applied Cognitive Psychology, 20 (5), pp. 689–696, 2006.

[19] Siegert R. J., McPherson K. M., Taylor W. J., Toward a cognitive-affective model of goal-setting in rehabilitation: is self-regulation theory a key step?, Disabil Rehabil, 26 (20), pp. 1175–1183, 2004.

[20] Ajzen I., The theory of planned behavior, Organ Behav Hum Dec Proc, 50, pp. 179–211, 1991.

[21] De Vries H., Dijkstra M., Kuhlman P., Self-efficacy: the third factor besides attitude and subjective norm as a predictor of behavioral intentions, Health Educ Res, 3, pp. 273–282, 1988.

[22] Ware J. E., Kosinski M., Dewey J. E., How to Score Version 2 of the SF-36 Health Survey, Lincoln, RI, Quality Metric Inc., 2000.

[23] Ware J. E. Jr, Kosinski M., Gandek B., et al., The factor structure of the SF-36 Health Survey in 10 countries: results from the IQOLA Project. International Quality of Life Assessment, J Clin Epidemiol, 51 (11), pp. 1159–65, 1998.

[24] Martin M., Kosinski M., Bjorner J. B., et al., Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale, Quality of Life Research, 16, pp. 647–660, 2007.

[25] Bjorner J., Ware J., Kosinski M., The potential synergy between cognitive models and modern psychometric models, Quality of Life Research, 12, pp. 261–274, 2003.

[26] Keller S. D., Ware J. E. Jr, Bentler P. M., et al., Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: results from the IQOLA Project. International Quality of Life Assessment, J Clin Epidemiol, 51 (11), pp. 1179–1188, 1998.

[27] Hayes V., Morris J., Wolfe C., Morgan M., The SF-36 health survey questionnaire: is it suitable for use with older adults?, Age Ageing, 24 (2), pp. 120–125, 1995.

[28] Lyons R. A., Wareham K., Lucas M., SF-36 scores vary by method of administration: implication for study design, J PublHlth Med, 21, pp. 41–45, 1999.

[29] Hambleton R. K, Patsula L., Increasing the Validity of Adapted Test: Myths to be Avoided and Guidelines for Improving Test Adaptation Practices, J Appl Testing Technology (JATT), 1 (1), pp. 1–30, 1999.

[30] Gimeno-Santos E., Frei A., Dobbels F., et al., Validity of instruments to measure physical activity may be questionable due to a lack of conceptual frameworks: a systematic review, Health and Quality of Life Outcomes, 9 (1), 86, 2011. http://www.hqlo.com/content/9/1/86

[31] Chmiel I., Górkiewicz M., Method for evaluating an individual inclination to guess Yes-or-No answers in case of a diversity of opinion at group of trustworthy responders, in: Bobrowski L., Burzykowski T., Doroszewski J., Enachescu C. (eds). 114-th ICB Seminar – VIII-th International Seminar: Statistics and Clinical Practice, Warszawa, pp. 59–61, 2011.

[32] Górkiewicz M., Chmiel I., Applying Rasch approach to comparative analysis of the of quality life measurements made with Polish version of the SF-36 questionnaire. in: Wybrane Determinanty Pielęgniarstwa, Część II, Sienkiewicz Z., Fidecki W., Wójcik G. (red.), Warszawski Uniwersytet Medyczny, Warszawa, pp. 128–136, 2010.

[33] Chmiel I., Determinants of quality of life following acute pancreatitis, Dysertacja doktorska, promotor: Antoni Czupryna. Uniwersytet Jagielloński, Wydział Nauk o Zdrowiu, Kraków, 2011.

[34] Scott K. M., Sarfati D., Tobias M. I., Haslett S. J., A challenge to the cross-cultural validity of the SF-36 health survey: factor structure in Maori, Pacific and New Zealand European ethnic groups, Soc Sci Med, 51 (11), pp. 1655–1664, 2000.

[35] Georgiadis K., Anthropological demography in Europe. Methodological lessons from a comparative study in Athens and London, Demographic Research, 17 (1), pp. 1–22, 2012. http://www.demographic-research.org/volumes/vol17/1/

[36] Bereziewicz W., Górkiewicz M., How much a priori in a posteriori: scientific recognition with use of the statistical methodology, Cogitatum, 2, pp. 1–9, 2012. on-line: http://filozof.uni.lodz.pl/knf/cogitatum/numer2/cog2bereziewicz.pdf

[37] Halonen K., Pettila V., Leppaniemi A., et al., Long-term health-related quality of life (HRQL) in survivors acute pancreatitis, Intensive Care Med, 29, pp. 782–786, 2003.

[38] Chmiel I., Górkiewicz M., Czupryna A., Brzostek T., Multiple comparisons procedures in analysis of health-related quality of life outcomes., in: Balcerar-Nicolau H., Bobrowski L., Doroszewski J., Kulikowski C. (eds). Lecture Notes of the VII-th ICB Seminar: Statistics and Clinical Practice, Warszawa, pp. 62–67, 2008.

[39] Chmiel I., Górkiewicz M., The Bootstrap and Multiple Comparisons Procedures as Remedy on Doubts about Correctness of ANOVA Results, Applied Medical Informatics, 30 (1), pp. 9–15, 2012.
http://ami.info.umfcluj.ro/index.php/AMI/article/view/352

[40] Górkiewicz M., Using propensity score with receiver operating characteristics (ROC) and bootstrap to evaluate effect size in observational studies, Biocybernetics and Biomedical Engineering, 29 (4), pp. 41–61, 2009.

[41] Chmiel I., Górkiewicz M., Czupryna A., Brzostek T., Vitality and feeling of happiness versus age, gender, physical functioning, and limitations in social role due to physical problems among convalescents after acute pancreatitis, Rocznik Naukowy, 19, Akademia Wychowania Fizycznego i Sportu w Gdańsku, Gdańsk, pp. 79–84, 2009.

[42] Chmiel I., Górkiewicz M., Czupryna A., Brzostek T., Age and gender as predictors of the physical ability among convalescents after acute pancreatitis., in: Fidecki W., Wysokiński M. (eds.) Selected problems of the aging population, Radomska Szkoła Wyższa, Radom, pp. 279–291, 2009.

[43] Chmiel I., Czupryna A., Górkiewicz M., Brzostek T., Health behaviours of covalescents after acute pancreatitis, in: Zdrowie, Kultura Zdrowotna, Edukacja. Czerwiński J., Demel M., Frołowicz T., et al. (eds.), Akademia Wychowania Fizycznego i Sportu w Gdańsku, Gdańsk, 2, pp. 145–150, 2008.

[44] Chmiel I., Czupryna A., Brzostek T., et al., Educational needs of patients after acute pancreatitis (preliminary report), Pielęgniarstwo XXI wieku, 28 (3), pp. 51–56, 2009.

[45] Chmiel I., Czupryna A., Górkiewicz M., Brzostek T., The causes of acute pancreatitis and the range of psychoeducational intervention for convalescents, Medical Studies, 11, pp. 51–56, 2008.

[46] Górkiewicz M, Chmiel I., Dutes of contemporary education of children and youth from perspective of health rehabilitation at convalescents after hard disease, in: Augustyn A., Bodanko A., Niestolik N. (red.n.) Dylematy współczesnego wychowania i kształcenia, Wyd. Akademii Humanistyczno-Ekonomicznej w Łodzi, pp. 113–118, Łódź, 2011.

[47] Dojka E., Górkiewicz M., Pająk A., Psychometric value of CES-D scale for the assessment of depression in Polish population, Psychiatria Polska, 37 (2), pp. 281–292, 2003.

[48] Peña-Sánchez J. N., Domagala A., Górkiewicz M., et al., Adapting a tool in Poland for the measurement of the physicians' career satisfaction, Problemy Medycyny Rodzinnej, 12 (1), pp. 58–65, 2011. http://pmr.org.pl/

[49] Wilczek-Rużyczka E., Czabanowska K., Walewska E., et al., Motivation to work fortifies attitude to motivating patients: Evidence from Leonardo da Vinci program on motivational skills training in health social care., in: Balcerar-Nicolau H., Bobrowski L., Doroszewski J., Kulikowski C. (eds). Lecture Notes of the VII-th ICB Seminar: Statistics and Clinical Practice, Warszawa, 113–119, 2008.

[50] Wilczek-Rużyczka E., Górkiewicz M., Decision variables of psychological model of nurse's to patients, in: Człowiek i jego decyzje, Kłosiński K.A., Biela A. (eds.), Wydawnictwo KUL, Lublin, 179–186, 2009.

[51] Górkiewicz M., Kreiner D.S., Gender Differenes in Creating False Memory under the DRM Paradigm, European Epi-Marker, 11 (2), pp. 6–12, 2007.

[52] Suhonen R., Leino-Kilpi H., Välimäki M., Development and psychometric properties of the Individualized Care Scale, Journal of Evaluation in Clinical Practice, 11 (1), pp. 7–20, 2005.

[53] Rothman M. L., Beltran P., Cappelleri J. C., et al., Patient-reported outcomes: conceptual issues, Value Health, 10 (2), pp. S66–S75, 2007.

[54] Bottomley A., Jones D., Claassens L., Patient-reported outcomes: assessment and current perspectives of the guidelines of the Food and Drug Administration and the reflection paper of the European Medicines Agency, Eur J Cancer, 45, pp. 347–353, 2009.

[55] Wiklund I., Assessment of patient-reported outcomes in clinical trials: the example of health-related quality of life, Fundamental & Clinical Pharmacology, 18 (3), pp. 351–363, 2004.
http://www.ncbi.nlm.nih.gov/pubmed/15147288

[56] Chmiel I., Górkiewicz M., The scope of acceptance by patients motherly and friendly style of nurse's supporting behaviour in palliative care, Problemy Pielęgniarstwa, 18 (4), pp. 11–17, 2010.

[57] Gniadek A., Kozicka M., Górkiewicz M., Unexpected, indirect and seeming associations between factors of the quality of life in elderly individuals, in: Wybrane Determinanty Pielęgniarstwa, Część II. Sienkiewicz Z., Fidecki W., Wójcik G. (eds.), Warszawski Uniwersytet Medyczny, Warszawa, pp. 118–127, 2010.

[58] SISA. On-line Calculators for Scientists. GraphPad Software, Inc., 2002–2012. http://www.quantitativeskills.com/sisa/

[59] StatSoft, Inc.How to Analysis Data with Low Quality or Small Samples, Nonparametric Statistics., in: StatSoft, Inc. Electronic Statistics Textbook, Tulsa, OK: StatSoft, 2012.
http://www.statsoft.com/textbook/nonparametric-statistics/

[60] Lowry R., The Friedman Test for 3 or More Correlated Samples., in: Concepts and Applications, 1998–2012. http://vassarstats.net/textbook/index.html

[61] Böckenholt U., Comparative Judgments as an Alternative to Ratings: Identifying the Scale Origin, Psychological Methods, 9 (4), pp. 453–465, 2004.