PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2023, TOM 66, NR 4, 151–167
DOI: 10.31648/przegldpsychologiczny.9467

Cronbach's alpha – what makes it really good?

Some advice for planning and criticizing psychological questionnaires

Tomasz Rak[1,2]
The Pontifical University of John Paul II in Krakow
https://orcid.org/0000-0002-3522-5176

Szymon Wrześniowski[2]
The Pontifical University of John Paul II in Krakow
https://orcid.org/0000-0001-5553-4016

Abstract

Whatever Cronbach's alpha measures – it's not internal consistency, commonly misunderstood in psychology as the average strength of relationships within questionnaire items. In this article, we explore the reasons why the understanding of alpha as internal consistency is particularly flawed, and focus on how alpha inflation works in a practical way. Using the simulation method, we determine the precise (common) influence of the number of respondents, the range of measurement (Likert) scales, the number of questions in the questionnaire and the average correlation of items on the alpha level. The results confirm alpha-level inflation due to a greater number of questions: alpha gets a satisfactory level even with minimal internal consistency if there are many questions in the questionnaire. We suggest that the reliability of weak psychological tools is overestimated because of presented rapid alpha inflation. Number of subjects and the range of the scale had no influence on alpha.

Keywords: reliability, alpha coefficient, alpha inflation, internal consistency, simulation

[1] Correspondence address: Tomasz Rak, tomasz.rak@o2.pl or tomasz.rak@upjp2.edu.pl.

[2] The authors certify that they, nor any member of their family, have NO affiliations with or involvement in any organization or entity with any financial, or non-financial interest in the subject matter or materials discussed in this manuscript, or any other interests or activities that might be seen as influencing or potential bias for the research.

In psychometrics, the reliability analysis allows to determine the precision (error) of the measurement (Revelle & Condon, 2019) of the research tool (usually a questionnaire) (Sijtsma & van der Ark, 2020). In other words, reliability is supposed to answer how well the test measures in a stable, repeatable way and what the magnitude of the possible measurement error is (Bajpai & Bajpai, 2014; Golafshani, 2003). According to the classical test theory, reliability is the magnitude of the correlation coefficient between the observed result and the true result, or correlation between parallel test scores (Metsämuuronen, 2022; Raykov & Marcoulides, 2011). In psychological literature, among many methods, by far the most popular measure of reliability is the Cronbach's alpha coefficient (see for example, Flake et al., 2017; McNeish, 2018; Ponterotto & Ruckdeschel, 2007; Taber, 2018) and it is on this factor that we want to focus.

The Cronbach's alpha measure has been repeatedly criticized in the literature in terms of its properties and interpretation (for example: Bonett & Wright, 2015; Dunn et al., 2014; Eisinga et al., 2013; Flora, 2020; Sijtsma & Pfadt, 2021). However, at this stage, we would like to mention that we do not intend to address the theoretical meanders of the phenomenon of measuring reliability (described more broadly by Borsboom & Mellenbergh, 2002 or Kane, 2013) or the quality of the alpha coefficient itself compared to other reliability coefficients (Anselmi et al., 2019; Trizano-Hermosilla & Alvarado, 2016). We will also try to avoid complex debates regarding the structure of formulas, so as not to generate problems with the reception of this content among psychology researchers who are not well-versed in mathematics (cf. Borsboom, 2006); we just want to make psychometrics simple again. This work is purely empirical in nature and our goal is to present a certain practical problem with the use of the alpha coefficient to those psychologists who simply use ready-made statistical packages in their work. The alpha coefficient should not be interpreted as the commonly understood *internal consistency*.

As indicated by numerous publications, *reliability* is often understood in psychology as the *internal consistency* of the results obtained within one psychological test (Kalkbrenner, 2021; Revelle & Condon, 2019). The alpha coefficient is frequently misinterpreted as a direct measure of *internal consistency* (Cho & Kim, 2015; Hayes & Coutts, 2020; Henson, 2001). It should, of course, be noted that in some cases *internal consistency* and reliability may be equal (Lucke, 2005; Ten Berge & Sočan, 2004), but this understanding in relation to the alpha coefficient is simply incorrect. At best, alpha's approaching the so-called greatest lower bound to the reliability, and in addition, even when assuming that items are tau-equivalent, GLB overestimate and may not express the reliability (Green & Yang, 2009; Sijtsma, 2009). Nevertheless, even leaving this fact aside, *internal consistency* itself, was and still is a subject of debate: how to understand it, what exactly is it, and what does it actually measure?

The very term *internal consistency* has so far been defined in various ways and the authors of works in the discussed area indicate that already at the definition level, its interpretation is the cause of much confusion (Bentler, 2009; Streiner, 2003; Tang et al., 2014). For example, the distinction between internal consistency, homogeneity, reliability, general factor saturation etc. is problematic, as these terms are often (and usually erroneously) used interchangeably.

Researchers also indicate that psychology generally identifies this *internal consistency* with the concept of whether a series of items (e.g., questionnaire questions) measures *more-or-less-the-same-thing* (McCrae et al., 2011; Sijtsma, 2009). This wording is also imprecise and, in consequence, confusing. However, it can be assumed, with a large degree of responsibility, that for psychologists, *internal consistency* often means that questionnaire items within one scale, are simply correlating strongly enough with each other when measuring a specific psychological property (see: Tang et al., 2014; Thigpen et al., 2017; Vaske et al., 2017). Therefore, psychology seems to confuse the reliability expressed by the alpha coefficient with the average correlation between the items in a questionnaire.

Sometimes, with regard to *internal consistency*, a one-dimensional structure of the psychological phenomenon is also assumed, i.e., that when tools such as factor analysis (usually PCA) or confirmatory analysis (CFA) are used, it will be possible to prove, that the studied structure is indivisible (see for example: Bentler, 2009; Gignac et al., 2007). This assumption is to some extent consistent with the idea of a high correlation of items within the assumed structure, and although different reduction methods will give slightly different results, when some items correlate more strongly, and some correlate less (cf. Jolliffe & Cadima, 2016; McDonald, 2013), it can still be assumed that we are dealing with the issue of relationships between items (questions) of a given questionnaire. Therefore, in order to demonstrate the effect we are interested in, and for the purposes of this article, we should define *internal consistency* as a commonly understood average correlation between items within a given psychological phenomenon. Therefore, our first area of interest in this article is the precise determination of the non-linear relationship between the mean correlation of the questionnaire items and the reliability measured by the alpha coefficient.

Certain inaccuracies or myths have also arisen in the area of reliability measurement using the Cronbach's alpha method. Some researchers have indicated that magnitude of alpha coefficients may depend on the number of items in the test, and greater number of questionnaire items usually inflates alpha (Cortina, 1993; DeVellis, 2006; Duhachek et al., 2005; Dunn et al., 2014). Indirectly, taking into account both the formula of the alpha coefficient itself, scaling the coefficient in relation to the number of items (Cortina, 1993; Thompson, 2003), or based on the prophetic Spearman-Brown formula (de Vet et al., 2017), it can be assumed that alpha coefficient inflation is a fact. However, the extent to which the reliability level is overestimated when the number of items measuring a given psychological construct is increased, remains unclear. Moreover, we believe that for psychologists who are not experts in mathematics, it can be extremely difficult to imagine how a certain curvilinear relationship will behave on a dual axis chart, based only on formulas. The second area of our interest is, therefore, a clear presentation of how the number of items truly inflates alpha, and a visual presentation of key overestimation thresholds.

Another issue of interest to us was the determination of the relationship between the alpha value and the length of the response scale that the subjects have at their disposal. So far, it has been suggested that the reliability of the obtained scales (latent dimensions of the questionnaire) improves with the length of the Likert scale used in the study (Leung, 2011; Preston & Colman, 2000; Taherdoost, 2019), although this does not necessarily apply to their *internal consistency*. This seems to be hardly possible in the context of the cited alpha calculation formulas which do not take into account the length of the scale used, because the alpha value should be based mainly on the relationships between the individual questions of the questionnaire. The variance of the results may be (and usually is under natural research conditions) different for the results collected using the same questions, but with different spans of the response scales. However, while the variance of the results itself plays an indirect role in the calculation of the alpha coefficient, it is certainly not clear how the length is the scale is to transfer into *internal consistency* (e.g., Chang, 1994; Matell & Jacoby, 1972). It was, therefore, also interesting for us to determine how large the differences in alpha values obtained would be between, for example, dichotomous scales (the subject gives a 'yes' or 'no' answer) and multi-point scales (seven-point Likert scales) for similar relationships between the elements (questions) of the questionnaire. It should be noted here that, some papers suggested, the alpha coefficient may not be suitable for determining the reliability of dichotomous scales (Barbaranelli et al., 2015; Pastore & Lombardi, 2014). This debate, however, is beyond our interest in this article. We will focus on internal consistency in the context of different scales lengths for which we obtained interesting results. These results will be discussed in the further part of the paper.

The structure of alpha formulas does not directly take into account the length of the scale on which the subject responds, nor the elements referring to the number of respondents (directly) appear there. Psychologists not professionally involved in statistics, do not usually ask themselves if there is any indirect relationship between the covariance matrix and the number of subjects (for example, formulas given in Brannick, 1995; Breckler, 1990 or Thall & Vail, 1990), but, for research purposes, psychologists are encouraged to increase the size of the sample on which a given measurement tool is validated (Chan, 2014; Zumbo & Chan, 2014). While it is certainly relevant for determining its structure (Tabachnick & Fidell, 2007) or its normalization (Guidroz et al., 2009; Macey & Eldridge, 2006), so far it has not been clear what effect the increase of the sample size actually has on both the alpha level and the aforementioned *internal consistency*.

Summarizing the above doubts, our article focuses on what impact the level of the average correlation between the items of a research tool, the number of items, the length of the response scale, and the number of subjects has on the Cronbach's alpha coefficient value.

Method

For the purposes of the study, the software in Delphi[3] was written. Its task was to generate a series of random datasets with different parameters based

[3] The code or pseudo-code of this software for generating fictitious data may be potentially dangerous (for example, allow for the fake of experimental research) and may only be made available after prior agreement with the authors.

on the Monte Carlo method (Dimov, 2008; Dunn & Shultis, 2011). Single datasets simulated the answers given by the respondents' during questionnaire surveys and differed as far as: the number of items (N_I), number of cases/subjects (N_N), range of response scales (N_L), and the average correlation between the items in the questionnaire (M_r). Each single dataset could contain from 2 (the minimum number to determine reliability) to 20 items (questions) of the questionnaire ($N_I = \{2,3,4…20\}$), and from 100 to 1000 cases (fictitious test subjects) prepared in increments of 100 ($N_N = \{100,200,300,...,1000\}$). Then, different ranges of response scales were established within the simulated research tool —- from a dichotomous scale (two-point scale, representing 'yes' vs. 'no' responses) to seven-point Likert scales, so the span of the scales ranged from 1–2 to 1–7 points ($N_L = \{2,3,4…7\}$). The last determined condition was the average correlation between the items of the questionnaire, measured with the Pearson's r coefficient, whereas its range was set from .10 to .90, in increments of .10 ($M_r = \{.10,.20,.30,...,.90\}$). The listed input conditions could therefore give 19 × 10 × 6 × 9 = 10,260 possible combinations of such simulated questionnaire tools (datasets). For each single combination of conditions, a series of 100 sets of questionnaires of a given type was generated, which gave a total of 1,026,000 single simulated questionnaire results with different parameters (or 10,260 series of datasets).

For the purposes of this paper, and for the sake of simplification, only positive matrices were analyzed (it is assumed that negative correlations indicate items that should be inverted prior to measuring reliability (as indirectly discussed by Bland & Altman, 1997)). In order to obtain the results as close to the real data as possible, the generator inserted the appropriate noise level. The algorithm generated such sets of responses in which the average correlation between items was within +/– .02 of the set threshold, for example, for the.50 threshold, it gave a real range of $.48 < r_M <= .52$. Therefore, despite the established generator step-values, in the end, a fairly diverse result of the average correlation was obtained, which reliably reflected the results obtained in real studies. Within 100 different sets with the same generator input parameters (one series of datasets), the correlation values between individual items also slightly differed. It was assumed that, for example, two items could correlate more strongly with each other, while with the third one, the correlation could be weaker than the assumed the threshold and so on. This, despite the finally similar mean correlations between two different sets, resulted in a large spread of different possible data with the desired final properties.

Combinations of the number of items (N_I), number of subjects (N_N), range of response scales (N_L), and the average correlation between items (M_r) were treated as independent variables. The dependent variable were the Cronbach's alpha values calculated for a given set. Raw datasets[4] were processed using an R script. For each dataset, an exact average correlation between the items

[4] Their size exceeds 12 GB, so they are not included as supplementary material, but may be available upon prior contact with the authors.

was calculated, followed by the Cronbach's alpha coefficient, which resulted in an output file containing 1,026,000 final observations (data points). These final data may have included such "cases" for which the calculated mean correlation was slightly weaker than the assumed correlation step threshold. This allowed for a greater concentration of data on the charts and the results development in a manner that would be closer to continuous functions than to step functions. The presentation uses averaged values for initial correlation steps for greater chart readability. The data file is available as supplementary material for this article[5].

Before proceeding with the analyses, an equivalence check of the research plan conditions was carried out. It was assumed that correctly generated data would have exactly the same average correlation levels in each combination of conditions due to the number of items (N_I), number of subjects (N_N), and the maximum range of the Likert scale (N_L). This result was, of course, obtained the sets can be considered equivalent, because the average correlation level in each of these conditions is practically the same (close to .51).

A certain difficulty in further analysis of the results was the indication of a "good" alpha measure. Despite increasingly restrictive recommendations, values above .80 and even above .70 are still acceptable in research papers (Taber, 2018). Hence, when presenting the results, we rather refer to the raw alpha value, allowing the reader to interpret the reliability independently on the basis of the coefficient value.

Results

The group size (N_N) had no effect on the alpha level in the set series. There was also no effect of the scale length (N_L) on the alpha level, expressed both in the form of a regression equation, Table 1, and the general linear model in Table 2 (p. 157). However, the number of questionnaire items (N_I) and the average correlation between them (M_r) turned out to be positively related to the alpha level. These three variables remained in a nonlinear relationship with each other and should be described in more detail.

The increase in the alpha level observed due to the value of the average correlation between items (M_r in the range from 0.1 to 0.9) is presented in the figure below (Figure 1, p. 158). This is a nonlinear relationship that for the generated data can be expressed as logarithm formula ($alpha$ = 1.02 + .21* LN(M_r)). A general conclusion can be drawn: with the increase in the mean correlation between items, the alpha value increased, but this increase was the strongest (steepest) up to the average correlation level close to 0.40. Then, the alpha exceeded the "acceptable" value of 0.80, and then the further increase in alpha was much smaller.

[5] A model size with this number of "cases" is not suitable for standard general linear modeling techniques in statistical packages. For the purposes of presenting significance models, a database with results averaged over one hundred sets of data was also made.

Table 1

Average Alpha Level Predictors for Simplified Dataset (Linear Regression Without Interaction)

predictor	*B (SE)*	β
(Intercept)	.45 (.00)	---**
N_L	.00 (.00)	.00
N_N	.00 (.00)	.01
N_I	.01 (.00)	.42**
M_r	.05 (.00)	.78**
F	9335.89**	
R^2	.784	

* $p < .05$,** $p < .01$; N_I – number of items / questions in the questionnaire, N_N – number of cases / subjects, N_L – maximum range of the Likert scale, M_r – mean correlation between the items of the questionnaire

Table 2

Average Alpha Level Predictors for Simplified Dataset (General Linear Model With Interactions)

Source of variation	*F*	*p*	η^2
N_L	.10	.992	< .001
N_N	.31	.973	< .001
N_I	974.78	< .001	.214
M_r	49735.05	< .001	.608
$N_L \times N_I$	< .01	1.000	< .001
$N_L \times N_I$	< .01	1.000	< .001
$N_I \times M_r$	364.24	< .001	.080
$N_L \times N_N$	.01	1.000	< .001
$N_L \times M_r$	.03	1.000	< .001
$N_N \times M_r$	.11	.999	< .001
$N_L \times N_N \times N_I$	< .01	1.000	< .001
$N_L \times N_I \times M_r$	< .01	1.000	< .001
$N_N \times N_I \times M_r$	< .01	1.000	< .001
$N_L \times N_N \times M_r$	< .01	1.000	< .001
$N_L \times N_N \times N_I \times M_r$	< .01	1.000	< .001

N_I – number of items / questions in the questionnaire, N_N – number of cases / subjects, N_L – maximum range of the Likert scale, M_r – mean correlation between the items of the questionnaire

Moreover, the increase in the alpha level followed the increase in the number of questionnaire items (Figure 2, p. 159). It is also a logarithmic relationship (*alpha* = .58+.12 × LN(N_I)). A fairly strong increase in alpha was observed up to the 5 items/questions of the questionnaire, where alpha was also close to 0.80. Both of these observed results, however, are only a simplified projection of the interaction of the mean correlation and the number of items onto a two-dimensional space. Let us examine this interaction in three dimensions.

For two items (questions) of the questionnaire (N_I = 2), the relationship between the mean correlation (M_r) and the alpha value was close to linear–the alpha value was almost equal to the average correlation level between items. However, with the increase in the number of items (questions) in the questionnaire, this relationship became increasingly curvilinear (Figure 3, p. 159, Table 3, p. 160) – with 20 questionnaire items (N_I = 20), very weak relationships between items (M_r ≈ .20-.30) were enough to obtain an alpha level usually reported as very good (.85). It is also worth noting that even with the minimum mean correlation (.10), along with the number of items, the mean alpha value increased to exceed the 0.70 alpha level with 20 items of the questionnaire. In other words, even if the questionnaire has very weak average relationships between individual questions, and thus a very bad *internal consistency*, a large number of questions in this questionnaire inflates alpha to a level where it seems to be already high or very high. This dependence can be expressed with a formula *alpha* = .0211+.1678 × M_r +.0531 × M_r –.0082 × M_r ^2–.0032 × M_r × N_I –.0011 × N_I ^2.

Figure 1

Cronbach's Alpha as the Average Result Obtained for the Questionnaire With an Assumed Level of Correlation Between Items

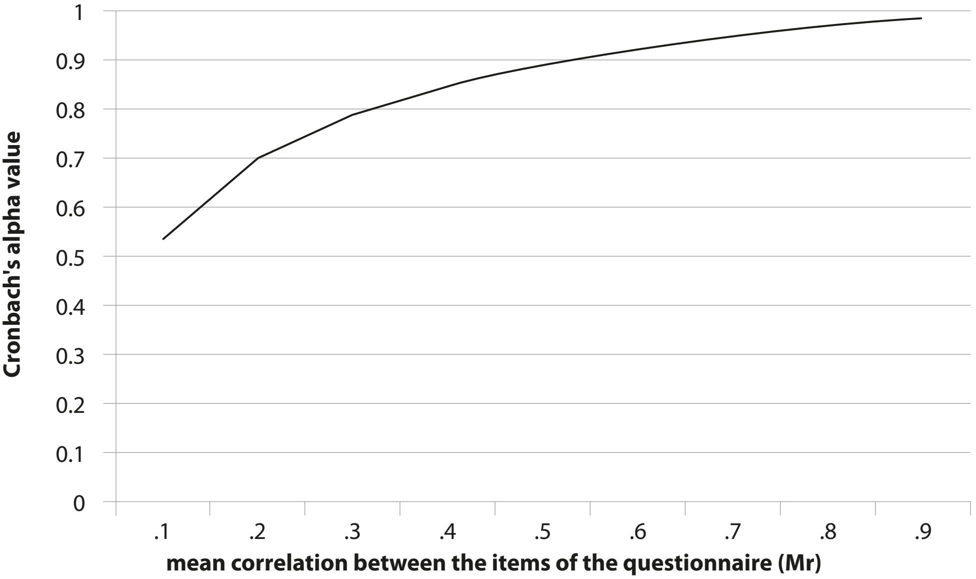

Figure 2

Cronbach's Alpha Value as the Average Result Obtained for the Questionnaire With an Assumed Number of Questions

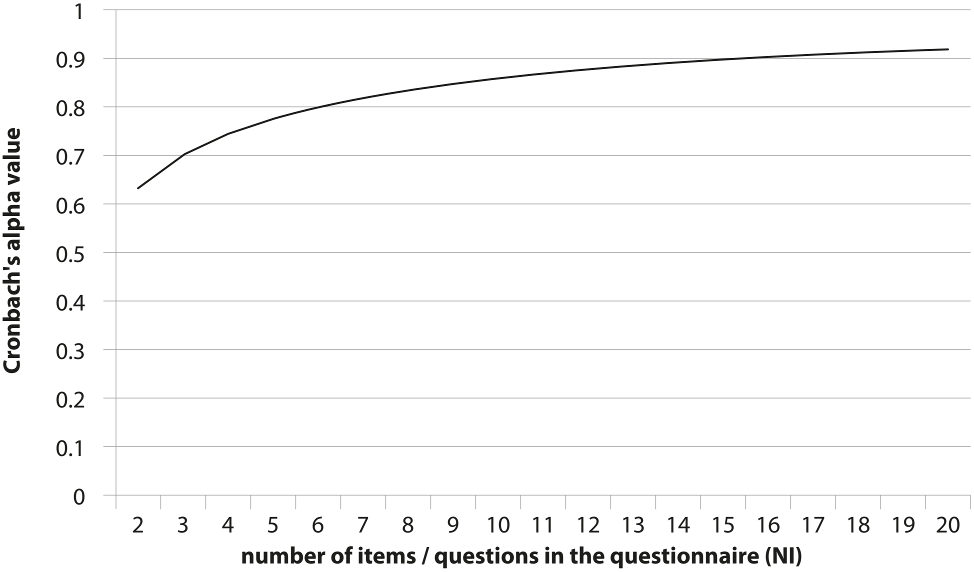

Figure 3

The Relationship Between the Mean Level of Correlation of the Questionnaire Components and the Mean Level of Cronbach's Alpha; Lines Represent Different Numbers of Questionnaire Items – From 2 Questions (Darkest Lower Lines) to 20 Questions (Brightest Higher Lines)

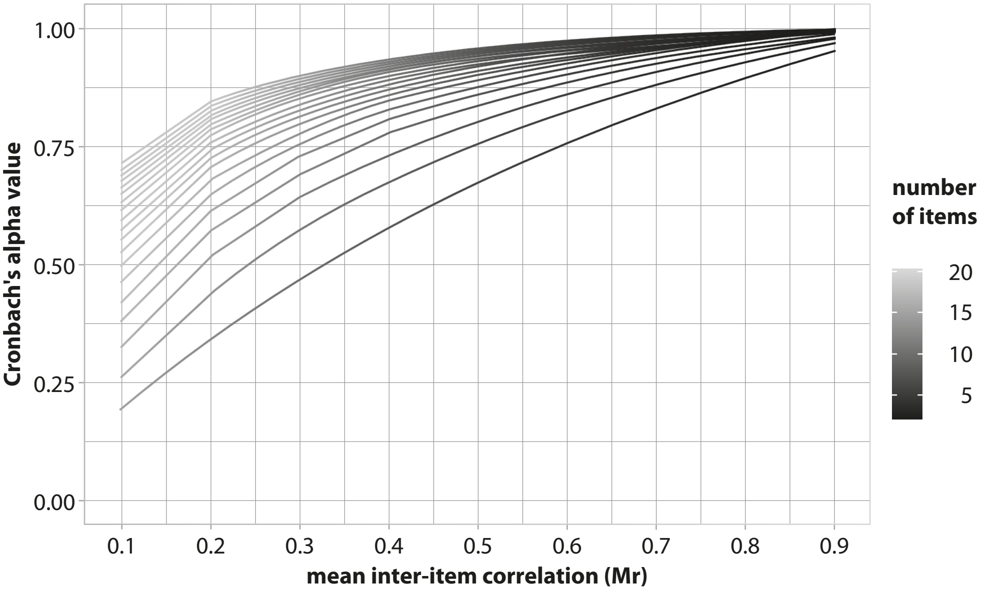

Table 3

The Relationship Between the Mean Level of Correlation of the Questionnaire Components (M_r) and the Number of Items / Questions in the Questionnaire (N_I) and the Mean Level of Cronbach's Alpha: Exact Results (Values in Table)

					M_r				
N_I	.10	.20	.30	.40	.50	.60	.70	.80	.90
I2	.192	.342	.468	.577	.672	.755	.828	.893	.951
I3	.265	.439	.571	.673	.755	.822	.878	.926	.967
I4	.326	.512	.640	.733	.804	.861	.906	.943	.975
I5	.378	.568	.690	.775	.837	.885	.923	.954	.980
I6	.422	.612	.727	.805	.861	.903	.935	.962	.983
I7	.461	.648	.757	.828	.878	.915	.944	.967	.986
I8	.494	.678	.781	.846	.892	.925	.951	.971	.987
I9	.524	.703	.800	.861	.903	.933	.956	.974	.989
I10	.550	.725	.817	.873	.912	.939	.960	.977	.990
I11	.574	.744	.831	.883	.919	.944	.964	.979	.991
I12	.595	.760	.842	.892	.925	.949	.967	.980	.991
I13	.614	.774	.853	.900	.931	.953	.969	.982	.992
I14	.632	.787	.862	.906	.935	.956	.971	.983	.993
I15	.648	.798	.870	.912	.939	.959	.973	.984	.993
I16	.662	.808	.877	.917	.943	.961	.975	.985	.994
I17	.676	.818	.883	.921	.946	.963	.976	.986	.994
I18	.688	.826	.889	.925	.949	.965	.977	.987	.994
I19	.700	.834	.894	.929	.951	.967	.979	.988	.995
I20	.710	.841	.899	.932	.954	.969	.980	.988	.995

Discussion and Conclusions

The analyses carried out demonstrated that the alpha level is influenced by the average correlation between questionnaire items and the number of items whereas, the number of subjects and the length of the response scale used has marginal importance (actually none).

The lack of impact of the group size on the alpha level can be indirectly predicted by taking into account the previously cited formulas describing the alpha coefficient (Cortina, 1993; de Vet et al., 2017; Thompson, 2003). It seems interesting that the results obtained from our study based on simulations of the subject response distributions are not consistent with the theoretical argument posed by Bujang et al. (2018), who postulate that in order to obtain a satisfactory reliability of the tested research tool, the necessary sample size. Without going into details, however, their theoretical argument concerns primarily small sample sizes ($N <= 10$), while we focused on sample sizes greater than $N = 100$[6]. A similar report for small samples can also be found in Šerbetar & Sedlar (2016) work.

The observed lack of influence of the response scale length is also related to the question of whether alpha is suitable for determining *internal consistency* for dichotomous scales (Barbaranelli et al., 2015; Pastore & Lombardi, 2014). Once again, it should be noted that we do not get an answer to the question of whether and how the scale length interacts with the *reliability* understood in psychometrics in numerous different manners. Nevertheless, in the presented results, whether the scale was dichotomous or had three, five or seven points, was of no relevance for the alpha level, because the basis of this indicator is (in simplification) the average correlation between positions. It can be pointed out (with a high degree of probability) that it is the relationships between items that determine the alpha coefficient level. From the methodological perspective, on the other hand, we can but wonder whether it is simply more difficult to obtain correlations between items in dichotomous scales than in larger scales. This, however, is not an issue this article addresses.

Naturally, the literature recommends the use of other measures, such as McDonald's Omega, but it is difficult to say if and when they will gain popularity, since they are difficult or impossible to use for standard functions of popular statistical packages (Hayes & Coutts, 2020). As pointed out above, our goal is not to compare alternatives to alpha. A reader who would like to choose the best of several dozen available coefficients should rather consult the works of Cho (2022) or Trizano-Hermosilla & Alvarado (2016).

It appears interesting that the relationship between the mean correlation and alpha level is so strongly curved by the number of items. From a practical point of view, reliability must refer to some psychological construct that is uniform in definition. This means that survey questions should concern a similar construct or a uniform topic – then we can talk about *reliability* in general. However, in the case where the questions are actually unrelated to each other (i.e. the respondents' answers are uncorrelated), a high level of reliability of a completely unconfirmable construct of an unknown nature can still be obtained. Based on the calculations performed, it can be said that with relationships that usually in psychology are often considered negligible, the alpha coefficient is at an acceptable level already with 20 items, and by increasing the number of questions, even with close-to-zero relationships within the questionnaire, it is possible to obtain almost perfect *alpha-reliability* for 90 items. An example of this is the adaptation of the popular in Poland Hobfoll's Cor-Evaluation questionnaire (Gruszczyńska, 2012), consisting of 90 questions which allows to obtain a general result – and this result is somehow *reliable*.

The fact that the number of items is so important for the alpha coefficient indicates its serious disadvantage in measuring the commonly understood

[6] Of course, it should be mentioned that large groups are usually recommended for validation of research tools (for example see Charter, 1999).

internal consistency. It may occur that the questionnaire has close-to-zero relationships within individual items, but it obtains a high alpha coefficient only due to the great number of questions. It should be added that the literature often recommends increasing the number of questions in the research questionnaire to increase the reliability of the measurement (hence the existence and common usage of the Spearman-Brown prophetic formula, which, by the way, was proposed to correct the reliability of long tests). Nevertheless, as we have shown, with (let us emphasize it again) negligible correlations between questionnaire items, a high reliability can still be obtained, which simply will just be overestimated and incorrect.

The relationship between the commonly understood *internal consistency* and the alpha result is in fact easy to determine: it is a curvilinear relationship which indicates that even with weak correlations within the measuring tool (around .30), it is possible to obtain fairly satisfactory reliability (close to alpha = .80), whereas with mean correlations around .50 – a very high one (close to alpha = .90). Naturally, our focus is not on when the correlations are "strong" and the reliability is "satisfactory" – we just want to point out that alpha is not a good measure of *internal consistency*, because it clearly overestimates it. It seems equally thought-provoking that, as a measure of reliability understood as *internal consistency,* alpha overestimates this very internal consistency. Perhaps the popularity of Cronbach's alpha method does not result from its ease-of-use factor or its availability in SPSS (as suggested by Borsboom, 2006)? We can also ask whether the popularity of alpha does not lie in how easy it is to show that the research tool created is "reliable" (whatever that means) and that it has a high alpha measure, even if the *internal consistency* is weak or actually zero/negligible.

Our results may contribute to a more accurate determination of the measurement reliability and a more conscious approach to the measurement error estimation. On the basis of the last attached chart or table, a fairly simple prediction can be made when planning and constructing an own measurement tool: knowing the average correlation level, it is easy to determine the number of questionnaire items needed to obtain the desired alpha level. It can also be used in the retrospective assessment of the quality of a given research tool, as knowing the alpha level and the number of items, it is possible to retrospectively determine the average level of correlation between questionnaire items (with some approximation, of course).

To sum up, the length of the scale and the number of subjects (over 100) do not have any particular significance in determining the level of reliability understood as *internal consistency*, and the alpha coefficient is clearly overestimated when the number of questions in the questionnaire is greater. This is a critical remark regarding alpha itself, but it can be used in planning the creation of measurement tools which, due to hard-to-observe latent features, may simply require increasing the number of questions to improve the parameters of the questionnaire (as suggested, for example, by Hoyt et al., 2006). On the other hand, we suggest caution when presenting the so-called good alpha measures for those research tools (psychological questionnaires) that just contain a lot of items. Therefore, as a solution to the problem, we want to propose that in practice, the average correlations with reliability measures should be reported. Low average correlation between items, high reliability and large number of questions may indicate this inflation error. To assess *internal consistency*, we recommend a rather mean level of correlation (which is in line with the recommendations of Cho & Kim, 2015). Given its popularity, we suggest simply approaching the alpha coefficient with a lot of caution for more than 8 items, using alternative measures, or simply understanding that measuring *internal consistency* is not exactly an alpha task. Finally, for tools composed of more than 20 items on a single scale, we recommend the use of a multi-faceted and more careful approach to testing their reliability than performing alpha analysis only.

In the spirit of open science, the authors of the article encourage reader to access open research data available in the digital repository under the address: https://tiny.pl/dttmk

References

Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology*, *10*, Article 2714. https://doi.org/10.3389/fpsyg.2019.02714

Bajpai, S., & Bajpai, R. (2014). Goodness of measurement: Reliability and validity. *International Journal of Medical Science and Public Health*, *3*(2), 112–115. https://doi.org/10.5455/ijmsph.2013.191120133

Barbaranelli, C., Lee, C. S., Vellone, E., & Riegel, B. (2015). The problem with Cronbach's alpha: comment on Sijtsma and van der Ark (2015). *Nursing Research*, *64*(2), 140–145. https://doi.org/10.1097/NNR.0000000000000079

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137–143. https://doi.org/10.1007/s11336-008-9100-1

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *British Medical Journal*, *314*(7080), Article 572. https://doi.org/10.1136/bmj.314.7080.572

Bonett, D. G., & Wright, T. A. (2015). Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior*, *36*(1), 3–15. https://doi.org/10.1002/job.1960

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*(6), 505–514. https://doi.org/10.1016/S0160-2896(02)00082-X

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, *16*(3), 201–213. https://doi.org/10.1002/job.4030160303

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin*, *107*(2), 260–273. https://doi.org/10.1037/0033-2909.107.2.260

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A review on sample size determination for Cronbach's alpha test: a simple guide for researchers. *The Malaysian Journal Of Medical Sciences: MJMS*, *25*(6), 85–99. https://doi.org/10.21315/mjms2018.25.6.9

Chan, E. K. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. In *Validity and validation in social, behavioral, and health sciences* (pp. 9–24). Springer.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement*, *18*(3), 205–215. https://doi.org/10.1177/014662169401800302

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559–566. https://doi.org/10.1007/978-3-319-07794-9_2

Cho, E. (2022). The accuracy of reliability coefficients: A reanalysis of existing simulations. *Psychological Methods*. Online first. https://doi.org/10.1037/met0000475

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–230. https://doi.org/10.1177/1094428114555994

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Erlbaum.

de Vet, H. C., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearman–Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, *85*, 45–49. https://doi.org/10.1016/j.jclinepi.2017.01.013

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*(11), 50–59. https://doi.org/10.1097/01.mlr.0000245426.10853.30

Dimov, I. T. (2008). *Monte Carlo methods for applied scientists*. World Scientific. https://doi.org/10.1142/9789812779892

Duhachek, A., Coughlan, A. T., & Iacobucci, D. (2005). Results on the standard error of the coefficient alpha index of reliability. *Marketing Science*, *24*(2), 294–301. https://doi.org/10.1287/mksc.1040.0097

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Dunn, W. L., & Shultis, J. K. (2011). *Exploring Monte Carlo methods*. Elsevier. https://doi.org/10.1016/B978-0-444-51575-9.00007-5

Eisinga, R., Grotenhuis, M. T., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, *58*(4), 637–642. https://doi.org/10.1007/s00038-012-0416-3

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, *3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Gignac, G. E., Bates, T. C., & Jang, K. L. (2007). Implications relevant to CFA model misfit, reliability, and the five-factor model as measured by the NEO-FFI. *Personality and Individual Differences*, *43*(5), 1051–1062. https://doi.org/10.1016/j.paid.2007.02.024

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, *8*(4), 597–607.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Gruszczyńska, E. (2012). Kwestionariusz Samooceny Zysków i Strat – polska adaptacja Cor-Evaluation Se Hobfolla i jej podstawowe właściwości psychometryczne [Profit and Loss Self-Assessment Questionnaire – Polish adaptation of Hobfoll's Cor-Evaluation Se and its basic psychometric properties]. In E. Bielawska-Batorowicz & B. Dudek (Eds.), *Teoria zachowania zasobow Stevana E. Hobfolla. Polskie doświadczenia* [*Stevan E. Hobfoll's theory of conservation of resources. Polish experience*]. Wydawnictwo Uniwersytetu Łódzkiego.

Guidroz, A. M., Yankelevich, M., Barger, P., Gillespie, M. A., & Zickar, M. J. (2009). Practical considerations for creating and using organizational survey norms: Lessons from two long-term projects. *Consulting Psychology Journal: Practice and Research*, *61*(2), 85–102. https://doi.org/10.1037/a0015969

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But… *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177–189. https://doi.org/10.1080/07481756.2002.12069034

Hoyt, W. T., Warbasse, R. E., & Chu, E. Y. (2006). Construct validation in counseling psychology research. *The Counseling Psychologist*, *34*(6), 769–805. https://doi.org/10.1177/0011000006287389

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), Article 20150202. https://doi.org/10.1098/rsta.2015.0202

Kalkbrenner, M. T. (2023). Alpha, Omega, and *H* internal consistency reliability estimates: Reviewing these options and when to use them. *Counseling Outcome Research and Evaluation*, *14*(1), 77–88. https://doi.org/10.1080/21501378.2021.1940118

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Leung, S. O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research*, *37*(4), 412–421. https://doi.org/10.1080/01488376.2011.580697

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, *1*(1), 98–107. https://doi.org/10.1037/1082-989X.1.1.98

Lucke, J. F. (2005). The α and the ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement*, *29*(1), 65–81. https://doi.org/10.1177/0146621604270882

Macey, W. H., & Eldridge, L. D. (2006). National norms versus consortium data: What do they tell us. In A. I. Kraut (Ed.), *Getting action from organizational surveys: New concepts, technologies, and applications* (pp. 352–376). Jossey-Bass.

Matell, M. S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology*, *56*(6), 506–509. https://doi.org/10.1037/h0033601

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28–50. https://doi.org/10.1177/1088868310366253

McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology Press. https://doi.org/10.4324/9781410601087

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Metsämuuronen, J. (2022). The effect of various simultaneous sources of mechanical error in the estimators of correlation causing deflation in reliability: Seeking the best options of correlation for deflation-corrected reliability. *Behaviormetrika*, *49*(1), 91–130. https://doi.org/10.1007/s41237-022-00158-y

Pastore, M., & Lombardi, L. (2014). The impact of faking on Cronbach's alpha for dichotomous and ordered rating scores. *Quality & Quantity*, *48*(3), 1191–1211. https://doi.org/10.1007/s11135-013-9829-1

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills*, *105*(3), 997–1014. https:/ /doi.org/10.2466/pms.105.3.997-1014

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge. https://doi.org/10.4324/9780203841624

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Šerbetar, I., & Sedlar, I. (2016). Assessing reliability of a multi-dimensional scale by coefficient alpha. *Journal of Elementary Education*, *9*(1/2), 189–196.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K. (2020). *Measurement Models for Psychological Attributes: Classical Test Theory, Factor Analysis, Item Response Theory, and Latent Class Models*. CRC Press. https://doi.org/10.1201/9780429112447-2

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*, *86*(4), 843–860. https://doi.org/10.1007/s11336-021-09789-8

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn & Bacon.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Taherdoost, H. (2022). What is the best response scale for survey and questionnaire design; review of different lengths of rating scale / attitude scale / Likert scale. *International Journal of Academic Research in Management*, *8*(1), 1–10.

Tang, W., Cui, Y., & Babenko, O. (2014). Internal consistency: Do we really know what it is and how to assess it. *Journal of Psychology and Behavioral Science*, *2*(2), 205–220.

Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*(4), 613–625. https://doi.org/10.1007/BF02289858

Thall, P. F., & Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, *46*(3), 657–671. https://doi.org/10.2307/2532086

Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, *54*(1), 123–138. https://doi.org/10.1111/psyp.12629

Thompson, B. (2002). *Score reliability: Contemporary thinking on reliability issues* (1st ed.). Sage Publications, Inc. https://doi.org/10.4135/9781412985789.n1

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: congeneric and asymmetrical measurements. *Frontiers in Psychology*, *7*, Article 769. https://doi.org/10.3389/fpsyg.2016.00769

Vaske, J. J., Beaman, J., & Sponarski, C. C. (2017). Rethinking internal consistency in Cronbach's alpha. *Leisure Sciences*, *39*(2), 163–173. https://doi.org/10.1080/01490400.2015.1127189

Vehkalahti, K., Puntanen, S., Tarkkonen, L. (2006). *Estimation of reliability: a better alternative for Cronbach's alpha*. Department of Mathematics and Statistics, University of Helsinki.

Zumbo, B. D., & Chan, E. K. (2014). *Validity and validation in social, behavioral, and health sciences*. Social Indicators Research Series, Vol. 54. Springer. https://doi.org/10.1007/978-3-319-07794-9