International Journal of Psychology

International Journal of Psychology, 2024
DOI: 10.1002/ijop.13249

Short and sweet: Comparing strategies for the reduction of questionnaires on self-criticism and social safeness while preserving construct validity

Marcello Passarelli[1], Laura Casetta[2], Luca Rizzi[2], Carlo Chiorri[3], Francesca Cassina[1], Sandro Voi[4], and Diego Rocco[5]

[1]National Research Council of Italy, Institute of Educational Technology, Genoa, Italy
[2]Associazione Centro di Psicologia e Psicoterapia Funzionale, Padova, Italy
[3]Dipartimento di Scienze della Formazione, University of Genoa, Genoa, Italy
[4]Life and Medical Sciences, College Lane Campus, University of Hertfordshire, Hatfield, UK
[5]Dipartimento di Psicologia dello Sviluppo e della Socializzazione, University of Padua, Padua, Italy

Measuring psychological traits with standardised questionnaires is an essential component of clinical practice and research; however, patients and participants risk fatigue from overly long and repetitive measures. When developing the short form of a questionnaire, the most widely used method for selecting an item subset uses factor analysis loadings to identify the items most closely related to the psychological construct being measured. However, this approach will tend to select highly correlated, homogeneous items and might therefore restrict the breadth of the construct examined. In this study, we will present Yarkoni's genetic algorithm for scale reduction and compare it with the classical scale reduction method. The algorithm will be applied to the shortening of three instruments for measuring self-compassion and social safeness (two unidimensional measures and a three-factor measure). We evaluated the shortened scales using correlation with long-form scores, internal reliability and the change in the correlations observed with other related constructs. Findings suggested that the classical method preserves internal reliability, but Yarkoni's genetic algorithm better maintained correlations with other constructs. An additional qualitative assessment of item content showed that the latter method led to a more heterogeneous selection of items, better preserving the full complexity of the constructs being measured.

Keywords: Scale abbreviation; Self-criticism; Social safeness; Construct validity; Scale reduction.

Evidence-based psychology is built on evidence; evidence, in turn, is built on data. Whether they be quantitative or qualitative, data—and our interpretation of them—are the foundation of our knowledge as social scientists. It follows that data quality is of paramount importance, and psychology has a strong tradition for the development of quantitative questionnaires that are both valid (they measure what they are supposed to measure; Nunnally, 1978) and reliable (they measure consistently over time; Cronbach, 1951).

Still, there is no such thing as a free lunch, and the process of data collection is no exception; collecting good data, both in quantity and quality, typically requires effort. When it comes to questionnaires, part of this toll is paid for by participants (or patients) themselves. Filling out a questionnaire might seem straightforward, but it demands participants' time and attention, often without any direct benefit to them. Unfortunately, one of the methods most widely used to develop questionnaires measuring latent psychological constructs—factor analysis—tends to produce long and sometimes repetitive sets of questions aimed at ensuring accuracy (Ziegler et al., 2014).

The lengthy questionnaires developed in this way have significant drawbacks. As participants become more tired of filling them in, the quality of their responses tends to decline. Many participants may drop out or avoid

Correspondence should be addressed to Marcello Passarelli, National Research Council of Italy, Institute of Educational Technology, Via de Marini, 6, 16149 Genoa, Italy. (E-mail: passarelli@itd.cnr.it).

starting altogether if the length is apparent from the outset (Bowling et al., 2021; Eisele et al., 2022; Galesic & Bosnjak, 2009; Iglesias & Torgerson, 2000). Furthermore, even if data quality is not impaired, it is inconsiderate (or even unethical) to burden participants with lengthy questionnaires, especially when the extra length provides little additional value and participants are not compensated. This could lead researchers to limit the number of questionnaires in a study due to concerns about the total number of items, but this, in turn, carries the risk of potentially omitting important psychological constructs from a study, limiting its depth and breadth. Lastly, in the clinical context, long measures take up valuable time, both to fill in the questionnaires and to score the results if scoring is not automatized.

Given the issues with lengthy questionnaires, there is a marked interest in shortening psychological measures (Kruyen et al., 2013). The aim is to reduce the number of items while preserving the core essence, validity and accuracy of the measure. Typically, this involves choosing questions that are most representative of the overall concept being measured or that are strongly linked to the construct in statistical terms, that is, items are selected because they highly correlate with the overall scale score or have the highest loadings on the target construct in a factor analysis (Kleka & Soroko, 2018; Kruyen et al., 2013). These methods present the advantage of maximising the shortened scale's reliability and the metrics used to evaluate the questionnaire when carrying out a confirmatory factor analysis—common key indicators of a quality questionnaire that may even facilitate publication. However, there is a substantial and insidious downside: these selection methods tend to favour the most content-homogeneous items, leading to a shortened form that, while still reliable, may measure a narrower construct than originally intended (Kleka & Soroko, 2018; Kruyen et al., 2013) and thus lose in validity. For example, consider the hypothetical correlation matrix in Table 1.

Assuming a sample size of 500, if we performed a parallel analysis (a standard method to investigate the dimensionality of a measure), we would observe that: (a) the downward curve of the eigenvalues flattens out from the second factor; (b) just one observed eigenvalue is larger than the simulated ones; and (c) the first-to-second eigenvalue ratio would be 5.01. Taken together, these results suggest that the optimal number of factors is one (Figure 1).

A single-factor exploratory factor analysis would explain 37% of variance, with all loadings larger than .50 (Table 2). If we adopted a confirmatory factor analysis approach and specified a single-factor model, we would obtain an acceptable model fit (CFI = .934, TLI = .919, RMSEA = .068 [.057, .079]). The internal consistency of the scale would be .87 (Cronbach's α), and the Cronbach's α-without-the item values suggest that all the items almost equally contribute to the scale's internal consistency.

Now, assume that one would like to include this scale in a large survey and that a shortened version of this scale would be preferable to limit administration time. Looking at the factor loadings in Table 2, it could be concluded that the first four items could be selected, since they have the highest factor loadings. Being this parameter, the correlation between the item and the factor score, they should—intuitively—be the most representative items. Unfortunately, this is not the case: if we carefully inspect the correlation matrix in Table 1, it is apparent that the first four items have between them a correlation (in the .50s) that is stronger than their correlation with all other items and than the correlation of all other items between them (in the .30s). It is likely these first four items are more similar one to another than with all the other items, and they are likely to tap into a narrower construct than the original. In other words, the shortened scale would not measure the same construct operationalised by the initial 12 items, but could measure a construct somewhat different. This

TABLE 1
Hypothetical correlation matrix

	i01	*i02*	*i03*	*i04*	*i05*	*i06*	*i07*	*i08*	*i09*	*i10*	*i11*	*i12*
i01	1.00											
i02	.51	1.00										
i03	.59	.54	1.00									
i04	.53	.51	.58	1.00								
i05	.38	.32	.35	.33	1.00							
i06	.38	.39	.32	.38	.35	1.00						
i07	.30	.36	.33	.37	.32	.30	1.00					
i08	.38	.31	.34	.38	.39	.33	.39	1.00				
i09	.35	.36	.40	.37	.32	.36	.38	.34	1.00			
i10	.39	.36	.32	.38	.38	.33	.34	.32	.31	1.00		
i11	.39	.35	.30	.30	.32	.39	.33	.36	.40	.32	1.00	
i12	.30	.37	.35	.39	.35	.31	.32	.36	.36	.35	.37	1.00

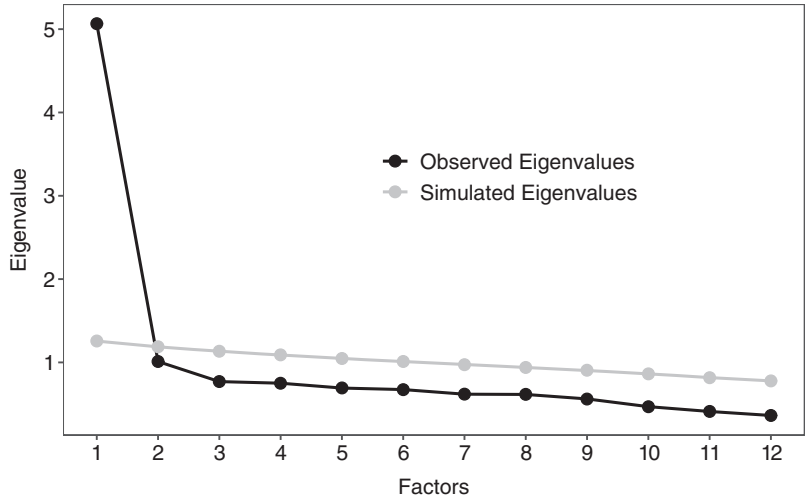

Figure 1. Scree-plot from the parallel analysis of data in Table 1 ($n = 500$).

would raise the issue of its validity while retaining acceptable unidimensionality and internal consistency, since the single factor would explain 55% of variance with loadings in the .70s and the Cronbach's α would be .83.

A practical example of this situation can be observed in the selection of items in the final version of the Interpersonal Exploitativeness Scale (Brunell et al., 2013). The authors defined "exploitativeness" as "the state, condition, quality, or degree of unfairly or cynically using another person or group for profit or advantage" (p. 2) and developed an initial item pool of 33 items. They administered the scale to 482 undergraduate students and, using EFA, obtained evidence for a one-factor solution. Given their goal of creating a short measure, they decided to retain only those items whose loading was larger than .70. The final six-item scale comprises statements that are very homogeneous in content (e.g., "It doesn't bother me to benefit at someone else's expense" and "I'm perfectly willing to profit at the expense of others"), that have a high average inter-item correlation (.54) and, consequently, a high Cronbach's α (.87). We are not here to judge that this procedure is correct or wrong, but looking at the item pool in Brunell et al. (2013)'s table 2, it is apparent that some content of the original pool has been lost. If we assume that this original item pool was a representative sample of the content domain of the construct, the final item pool does not seem to retain this property, since its conceptual breadth is narrower.

In practical settings, any shortened questionnaire may lack relevant operationalisations of the construct that is meant to measure. To avoid this issue, it is worthwhile to explore alternative methods for scale reduction. This paper will present and illustrate an alternative method, comparing it with the traditional reduction approach and showing how it circumvents the issues outlined.

TABLE 2
Factor loadings (λs) for the exploratory (EFA) and confirmatory factor analysis (CFA) and alpha without the item (α w/o) for data from Table 1 ($n = 500$)

Item	*EFA* λ	*CFA* λ	α *w/o*
i01	.69	.70	.86
i02	.67	.68	.86
i03	.68	.70	.86
i04	.69	.71	.86
i05	.56	.55	.87
i06	.57	.56	.87
i07	.55	.54	.87
i08	.58	.57	.87
i09	.59	.58	.87
i10	.56	.56	.87
i11	.56	.55	.87
i12	.57	.56	.87

Note: All CFA parameter estimates are significant at $p < .001$.

The technique we will consider, proposed by Yarkoni (2010), uses a Genetic Algorithm. This algorithm simulates natural evolution by randomly generating different short versions of a scale, including different sets of items. These sets of items are evaluated according to a specified criterion. The sets of items that are most "fit" are retained, then randomly altered into slightly different sets (akin to natural "mutations" of an organism), and then again evaluated. The process continues until the algorithm converges to a solution that, according to the specified fitness criterion, appears to be optimal and further alterations of the item sets lead to no improvement. In our context, the fitness function aims to maximise the variance explained by a linear combination of items while penalising the number of items used. Although this method may still favour items with high item-total correlation, it typically avoids selecting pairs

of highly correlated—and therefore homogeneous or even redundant—items.

In the following sections, we will use the genetic algorithm to shorten three instruments with different features and compare its effectiveness with that of the classical reduction method. To ensure that the context of the study will be relevant to both researchers and practitioners, we will test and compare scale reduction techniques in the context of a study on self-compassion and perceived social safety, in which long measures were administered to many participants. Additionally, in the tutorial section we will also apply the genetic algorithm to the shortening of the hypothetical scale simulated in Table 1, to compare the two different abbreviation strategies in a case in which the ground truth is known.

METHODS

Participants and procedure

The scale reduction methods will be compared using a dataset collected for a study on self-compassion designed with clinical psychologists. When designing the study, clinicians themselves expressed the need for shorter scales, as they found that the questionnaires were too cumbersome for their clinical setting.

Data were collected through an online survey that included a battery of five questionnaires. The study involved 733 total participants. However, we excluded from the dataset participants with more than 10% of missing responses, bringing the total of cases used in the analyses to 409 (102 men, 288 women, 19 undisclosed/other; age 40.83 ± 12.43 years).

Materials

Of the five questionnaires included in the study, three of them, which present radically different features, will be used to test and compare the genetic algorithm and the classic scale reduction method, while the other two questionnaires will be used to evaluate the obtained reduced scales (see Evaluation metrics section). The three questionnaires selected for scale reduction are the following:

Early Memories of Warmth and Safeness Scale (EMWSS)

The EMWSS (Richter et al., 2009) is a 21-item scale for measuring personal emotional memories of feeling warm, safe, and cared for in childhood. Example items include "I felt cared about," "I felt appreciated the way I was" and "I felt part of those around me." The response scale for each item ranges from 0 ("No, never") to 4 ("Yes, most of the time"). The scale is unidimensional and represents a prime candidate for scale reduction, as its items are highly redundant (e.g., item 10, "I could easily be soothed by people close to me when I was unhappy," item 17, "I knew I could rely on people close to me to console me when I was upset," and item 20, "I knew that I could count on help from people close to me when I was unhappy", have clear overlaps). Item redundancy results in an especially high Cronbach's α (.91 in the validation study, .97 in our sample).

Social Safeness and Pleasure Scale (SSPS)

The SSPS (Gilbert et al., 2009) is an 11-item self-report measure of social safeness. The items relate to feelings of belonging, reassurance and warmth from others (e.g., "I feel easily soothed by those around me"; "I feel a sense of warmth in my relationships with people"). The response scale for each item ranges from 1 ("Almost never") to 5 ("Almost all the time"). This scale, while considerably shorter than the EMWSS, could still be considered slightly redundant (e.g., item 3, "I feel connected to others," item 4, "I feel part of something greater than myself" and item 7, "I feel a sense of belonging"). The scale has an internal consistency of Cronbach's α = .91 in the validation study, and .92 in our sample.

Forms of Self-Criticising/Attacking and Self-Reassuring Scale (FSCRS)

The FSCRS (Gilbert et al., 2004) is a 22-item scale measuring self-criticism and self-reassurance in the face of adversity. The response scale for each item ranges from 0 ("Not at all like me") to 5 ("Extremely like me"). The scale has three distinct subscales: Inadequate-self (e.g., "When things go wrong for me, I am easily disappointed with myself"), Hated-self (e.g., "I have a sense of disgust with myself"), and Reassured-self (e.g., "I am gentle and supportive with myself"). This scale was selected to provide an example of scale reduction of a multidimensional scale. While it is the longest questionnaire being shortened, since it measures three related constructs, it is the least redundant of the three. Internal consistency for the three factors was .90, .86 and .86, respectively, for the validation study, and .91, .82 and .90 for our sample.

The following two questionnaires will be used to test the quality of the reduced scales as described in the Evaluation metrics section.

Self-Compassion Scale (SCS)

The SCS (Neff, 2003) is a 26-item questionnaire investigating self-compassion, defined as "the ability to hold one's feelings of suffering with a sense of warmth, connection, and concern" (Neff & McGehee, 2010, p. 226). The response scale for each item ranges from 1 ("Rarely or never") to 7 ("Almost always"). The model

includes six correlated facets measuring different aspects of self-compassion: self-kindness (e.g., "I try to be understanding and patient towards those aspects of my personality I don't like," $\alpha = .78$ in the validation sample, .92 in this study), self-judgement (e.g., "When I see aspects of myself that I don't like, I get down on myself," $\alpha = .77$ in the validation sample, .88 in this study), common humanity (e.g., "I try to see my failings as part of the human condition," $\alpha = .80$ in the validation sample, .79 in this study), isolation (e.g., "When I fail at something that's important to me I tend to feel alone in my failure," $\alpha = .79$ in the validation sample, .83), mindfulness (e.g., "When I fail at something important to me I try to keep things in perspective," $\alpha = .75$ in the validation sample, .81 in this study) and overidentification (e.g., "When I'm feeling down I tend to obsess and fixate on everything that's wrong," $\alpha = .81$ in the validation sample, .80 in this study). The relatively low number of items compared to the number of subscales, as well as the low internal consistency—suggesting low item redundancy—do not make this scale a good candidate for shortening.

Fears of Compassion (FOC)

The FOC (Gilbert et al., 2011) is a 37-item questionnaire measuring fears, blocks and internal resistances to compassion. The response scale for each item ranges from 0 ("Don't agree at all") to 4 ("Completely agree"). It includes three subscales: fears about giving compassion to others (e.g., "Being too compassionate makes people soft and easy to take advantage of," $\alpha = .78$ in the validation sample, .83 in this study), fears about receiving compassion from others (e.g., "When people are kind and compassionate towards me I feel anxious or embarrassed," $\alpha = .87$ in the validation sample, .89 in this study) and fears of self-compassion (e.g., "I fear that if I become too compassionate to myself I will lose my self-criticism and my flaws will show," $\alpha = .85$ in the validation sample, .93 in this study). Again, the internal consistency, especially for the first subscale, suggests low item redundancy.

Scale Reduction Strategy

The dataset has been randomly split into a training dataset ($n = 283$) and a test dataset ($n = 119$). For each of the EWMSS, SSPS and FSCRS, scale reduction has been performed using the training dataset in the following way:

(1) Classical method (items with the highest factor loading on the single factor). We used Spearman-Brown's prophecy (Brown, 1910; Spearman, 1910) to choose a desired length for the questionnaire.
Since Cronbach's α depends on both the average inter-item correlation and the number of items, Spearman-Brown's prophecy can be used to compute the expected α when reducing the number of items:

$$\alpha_{new} = \frac{R\alpha}{1 + R\alpha}$$

where, α_{new} is the Cronbach's α after removing/adding the items, and R is the ratio between the number of items in the modified and original version. Rearranging the formula, one can compute the number of items (p) to remove/add from the original $p_{original}$ item pool to achieve a desired level of α_{new}:

$$p = \frac{p_{original} \times \alpha_{new}(1 - \alpha)}{\alpha\left(1 - \alpha_{new}\right)} - p_{original}$$

In this study, we used Spearman-Brown's prophecy to find the number of items for each scale that would be expected to have a Cronbach's α of .75, a commonly used threshold. This result determined how many items with the highest factor loading should have been included in the shortened scale;

(2) Genetic method: we performed scale reduction using the Genetic Algorithm as implemented in the GAabbreviate R package (Sahdra et al., 2016). We set the item cost penalty to .001 (very low), using instead the MaxItems argument to ensure we would obtain a reduced scale of the same length as with the previous method. This constraint is not necessary when using the genetic method, as a higher cost penalty could be used to let the algorithm itself find the optimal scale length. However, we wanted the scale length to be the same as that obtained with the classical method to ensure a fair comparison.

Therefore, for each questionnaire, we obtained two separate versions of a shortened scale, one with the classical method and one with the Genetic Algorithm. The scores for the two shortened versions were then computed, in addition to the scores for the original full scales.

In the case of the FSCRS, which is the scale with three separate factors, the above procedure was performed twice. First, we did it considering the instrument (correctly) as measuring three distinct, yet correlated, psychological constructs. Then, we performed the reduction procedure on the full scale, treating it as if it were a unidimensional scale. The latter reduction is purposefully incorrect, as items belong to separate facets of the same construct, but it will be used to evaluate how much the reduction methods tend to narrow the conceptual breadth of the construct being measured in the scale reduction process (in other words, how much they tend to select items belonging to only one facet when starting item heterogeneity is high).

Evaluation metrics

The shortened scale versions were compared using the following metrics:

(1) Correlation between the shortened scale score and the original scale score. A high correlation indicated that, despite shortening the questionnaire, the score obtained by participants was roughly the same. The closer to 1, the better;
(2) Difference between the actual Cronbach's α of the shortened scale and the expected one from the Spearman-Brown prophecy. If the Cronbach's α of the shortened scale was equal to or higher than the expected, we considered the scale as sufficiently reliable;
(3) Difference in correlations with other measures. Each scale score was correlated with the score of all other measures included in the survey battery (including the SCS and the FOC). Ideally, the scores of a shortened scale would have the same correlations with other constructs as the full scale's scores. This would suggest that the validity of the construct based on the reduced scale was the same as that on the original scale. For each (sub)scale, we computed the change in correlation as the standard deviation of difference between the correlations observed using the original instrument, and correlations observed using the shortened (sub)scale candidate. The change in correlations was also tested for statistical significance using a test for the difference between dependent correlations (Meng et al., 1992), correcting for multiple comparisons using Benjamini-Hochberg's correction (Benjamini & Hochberg, 1995). The lower the change in correlations, the better.

All metrics were computed using, separately, both the training and the test dataset. Shortened scales were expected to perform worse on the test dataset, as it was not the one used for the scale reduction process, but performance on the test dataset was essential as it is an indicator of the generalisability of results.

RESULTS

Spearman-Brown's prophecy formula suggested a length of two items for the EMWSS and three for the SSPS. For the FSCRS, the suggested length was 4 when considered (incorrectly) as unidimensional, and 11 when considering the dimensions separately (four items for hatred, three for inadequacy, four for self-kindness; see Data S1 for analyses on the scales' factorial structures).

The two scale reduction methods always suggested to retain different items, except for the self-kindness scale of the FSCRS, in which the classical and genetic methods converged to the same solution. Correlations between the shortened questionnaires and the original scales are reported in Table 3. Overall, the genetic algorithm seemed to be the method most effective at replicating original scale scores.

Cronbach's α is reported in Table 4, together with inter-item correlations (mean ± standard deviation).

TABLE 3
Correlations between shortened versions of the scales and the original questionnaires' scores

Questionnaire	*Classical method (training)*	*Genetic method (training)*	*Classical method (test)*	*Genetic method (test)*
EMWSS	.91	**.94**	.90	**.93**
SSPS	.93	**.95**	**.95**	.95
FSCRS (unidimensional)	.90	**.94**	.89	**.94**
FSCRS (hatred)	.97	**.99**	.98	**.99**
FSCRS (inadequacy)	.93	**.95**	.92	**.94**
FSCRS (self-kindness)	**.95**		**.96**	

Note: Numbers in bold indicate the best-performing method.

TABLE 4
Cronbach's α for the shortened scales, together with inter-item correlations (mean ± standard deviation)

Questionnaire	*Full scale (training)*	*Classical method (training)*	*Genetic method (training)*	*Full scale (test)*	*Classical method (test)*	*Genetic method (test)*
EMWSS	.97 (.63 ± .07)	.92 (.86)	.83 (.71)	.97 (.63 ± .08)	.92 (.86)	.81 (.68)
SSPS	.92 (.51 ± .11)	.83 (.62 ± .03)	.81 (.59 ± .00)	.93 (.56 ± .14)	.85 (.66 ± .06)	.84 (.64 ± .05)
FSCRS (unidimensional)	.94 (.06 ± .44)	.87 (.62 ± .05)	.77 (.00 ± .51)	.95 (.05 ± .47)	.86 (.60 ± .06)	.77 (−.04 ± .51)
FSCRS (hatred)	.81 (.47 ± .10)	.80 (.50 ± .11)	.72 (.40 ± .05)	.81 (.46 ± .10)	.77 (.46 ± .11)	.74 (.41 ± .08)
FSCRS (inadequacy)	.91 (.53 ± .08)	.84 (.62 ± .02)	.82 (.60 ± .07)	.91 (.54 ± .09)	.84 (.64 ± .01)	.83 (.62 ± .04)
FSCRS (self-kindness)	.88 (.49 ± .07)	.82 (.53 ± .08)		.92 (.58 ± .06)	.85 (.59 ± .08)	

Note: Shortened versions on the EMWSS have no standard deviation for inter-item correlation, since they comprise only two items. Metrics for the full scales are reported for comparison. Numbers in red indicate potentially critical internal consistency.

TABLE 5
Standard deviation of differences between correlations observed using the full scales and correlations observed using the shortened scales

Questionnaire	*Classical method (training)*	*Genetic method (training)*	*Classical method (test)*	*Genetic method (test)*
EMWSS	.050 (29%)	**.033 (7%)**	.092 (64%)	**.036 (0%)**
SSPS	.036 (21%)	.031 (29%)	.044 (21%)	**.026 (0%)**
FSCRS (unidimensional)	.048 (45%)	**.020 (0%)**	.096 (64%)	**.038 (18%)**
FSCRS (hatred)	**.006 (0%)**	.017 (27%)	.020 (0%)	**.014 (0%)**
FSCRS (inadequacy)	.054 (64%)	**.030 (27%)**	.050 (9%)	**.033 (9%)**
FSCRS (self-kindness)	.022 (27%)		.024 (0%)	

Note: Numbers in bold indicate the best-performing method.

Some of the scales shortened using the classical method still presented an $\alpha > .90$, suggesting that there could be residual item redundancy—e.g., item correlations larger than .70. The genetic algorithm, on the other hand, always obtained αs between .70 and .90, suggesting good internal consistency with lower levels of redundancy.

Lastly, Table 5 includes change in correlations when using the shortened scales, as compared to correlations obtained using the original, full scales. Percentages indicate how many of the correlations examined changed in a statistically significant way when using the shortened scales. The correlations themselves are reported in Data S1.

On the training dataset, performance seems to be slightly in favour of the genetic algorithm, with three out of six comparisons with lower deviation for correlations, and less correlations changing significantly. For the remaining three comparisons, one is in favour of the classical method, one is a tie (since the shortened scales are equivalent) and one has an ambiguous result (less change in correlations for the genetic algorithm, but more changes are statistically significant).

On the testing dataset, on the other hand, the genetic algorithm achieves better performance for all subscales examined (and one tie for the scale that converged with the classical method). This result suggests that the genetic algorithm creates shortened scales that better generalise to novel contexts, while the classical method tends to create shortened scales that perform optimally only on the dataset used for the shortening itself.

Notably, if we examine the case in which the FSCRS was incorrectly treated as if it were a unidimensional scale, we can observe that the classical method selected four items all belonging to the same subscale (Inadequacy). In effect, the scale shortened using the classical method has become a measure of sense of inadequacy, rather than self-kindness, unduly narrowing the construct being measured. In contrast, the Genetic Algorithm retained items from all subscales, obtaining a more complete coverage of the construct.

TUTORIAL

The application of the Genetic Algorithm is straightforward, as it has been implemented in the GAabbreviate R package (Sahdra et al., 2016). To perform the analysis, we used the GAabbreviate() function, where the first argument—*items*—is a matrix or data frame containing item scores for the full scale, and the second argument—*scales*—is a matrix or data frame containing the (sub)scale scores. For a unidimensional scale, the latter argument is a single vector.

In this study, we additionally set the number of items for the shortened scale using the *maxItems* argument and a very low item penalty (*itemCost* argument). As a result, our function calls looked like the following:

```
GAabbreviate(items=df[,1:10], scales=df[,11:12],
maxItems=4, itemCost=.001)
```

Alternatively, one may opt to let the algorithm itself choose the length of the shortened scale, by omitting the *maxItems* argument and setting the *itemCost* to a desired value (the default is .05):

```
GAabbreviate(items=df[,1:10], scales=df[,11:12],
itemCost=.05)
```

There is no optimal, agreed-upon value for the *itemCost* argument, and it should be determined by trial and error according to the characteristics of the specific scale. On our dataset, the default of .05 led to scales that were shorter than what we specified, and worse-performing in terms of correlations with the original scale and with the other instruments.

Importantly, as the genetic algorithm relies on random number generation, for the purposes of reproducibility the random seed should be set prior to running the algorithm using the *set.seed* function (e.g., *set.seed*(1)), or by using the seed argument of the function itself (e.g., *seed* = 1). It is also possible to increase the number of iterations using the maxiter argument, or the size of the populations in the evolutionary simulation using the argument popSize; as the algorithm is relatively fast, it is possible to increase these values beyond their default values of 100 and 50

with little cost in computation time. Lastly, the optional argument minR can be used to set a minimum item-total correlation for retaining an item.

The GAabbreviate() function returns an object of class GAabbreviate; we are interested in the $measure object it contains. For example, if we simulate 500 observations from the correlation matrix reported in the hypothetical example in Table 1 and then run the GAabbreviate() function with the arguments maxItems = 4 and seed = 1, we get the following output:

```
$items
 x2  x4  x5 x11
  2   4   5  11

$nItems
[1] 4

$key
      [,1]
[1,]    1
[2,]    1
[3,]    1
[4,]    1

$nScaleItems
[1] 4

$alpha
            Scale1
alpha 0.6721004

$ccTraining
[1] 0.9173403

$ccValidation
[1] 0.9200538
```

In this case, the algorithm is returning a four-item abbreviated version (indicated by $nItems) and is suggesting retaining the 2nd, 4th, 5th and 11th items ($items object) of the input item matrix. Importantly, this output refers to items by their position in the input item matrix, disregarding column names.[1] This can lead to some confusion if the item numbers of a scale don't correspond to their position in the matrix. The output also includes the obtained Cronbach's α for the reduced scale, here somewhat low (.67). The last two elements of the output are the correlations with the full-scale scores computed during the training and validation steps of the algorithm, respectively. It is possible to turn off the cross-validation performed by the function by setting the crossVal argument to FALSE. This would perform the abbreviation with the full input dataset, at the cost of losing a valuable measure of generalizability of the results.

In this simulated example, items 1–4 of the original scale correlate between .51 and .58, while the rest of the matrix has correlations between .30 and .40. This simulates a case in which a subset of items presents redundant wording and/or measures a subfactor of the construct being measured. We can observe that the genetic algorithm selected two of these items (item 2 and 4), while the classical method would result in an abbreviated form that would retain those four items exclusively. In this case, abbreviation using the classical method would result in a short form that measures, in practice, only part of the construct it would have been meant to measure.

DISCUSSION AND CONCLUSIONS

In this study, we shortened three measures related to self-compassion and social safeness using the classical and the genetic methods, and we compared the different versions of the shortened scales with each other. Overall, while the classical method led to short measures with higher internal consistency, the genetic method seemed to be more capable of preserving the conceptual breadth of the construct being measured. This is suggested by its ability to better maintain the correlations between the scale being shortened and scores on other closely related scales, a feature especially important in research settings. Furthermore, qualitative evaluation of the items in the shortened scales confirmed that the items retained by the genetic method were more heterogeneous in their content (see Tables 6 and 7), especially for the EMWSS and the unidimensional FSCRS, thus mapping a wider extent of the examined construct. The simulated example reported in the Tutorial section further highlights how the classical abbreviation method tends to select the items most correlated between each other, while the genetic algorithm preserves item diversity.

Overall, we highly recommend using the genetic algorithm for shortening scales. In our study, it managed to reduce the EMWSS by 90%, the SSPS by 73% and the FSCRS by 50%. This considerable reduction achieved high correlations with the original scales—all above .93—and presented better performance than the classical method at preserving correlations with related constructs (preserving, in the worst case, 71% of the correlations against the 36% of the classical method).

However, it should be kept in mind that this method is appropriate only in those cases in which the structure of the original, full-length scale is essentially unidimensional. Testing the unidimensionality of a scale can be a complex task (see, e.g., Raykov & Pohl, 2013; Reise et al., 2013; Rodriguez et al., 2016), but common

[1] While performing the analysis, we became aware of a bug that interests the current version of the package (1.3). The GAabbreviate() function usually strips the input item matrix of column names when handling missing data. However, if the input matrix has no missing data, column names will be retained and cause the function to throw as error ("I am stopping because of improper input. See above for a list of bad item(s)"). It's possible to avoid this error by explicitly stripping the input matrix of column names, for example using colnames(inputmatrix) <- NULL.

TABLE 6
Shortened versions of the EMWSS and SSPS

EMWSS-short (classic method)	*EMWSS-short (genetic method)*	*SSPS-short (classic method)*	*SSPS-short (genetic method)*
I knew I could rely on people close to me to console me when I was upset	I knew I could rely on people close to me to console me when I was upset	I feel secure and wanted	I feel connected to others
I knew that I could count on help from people close to me when I was unhappy	I felt at ease	I feel a sense of belonging	I feel secure and wanted
		I feel understood by people	I feel understood by people

TABLE 7
Shortened versions of the FSCRS

FSCRS-short unidimensional (incorrect—classic)	*FSCRS-short unidimensional (incorrect—genetic)*	*FSCRS-short 3-factor (correct—classic)*	*FSCRS-short 3-factor (correct—genetic)*
I am easily disappointed with myself	I feel beaten down by own self-critical thoughts	(Hatred) I have become so angry with myself that I want to hurt or injure myself	(Hatred) I have become so angry with myself that I want to hurt or injure myself
I find it difficult to control my anger and frustration at myself	I have become so angry with myself that I want to hurt or injure myself	(Hatred) I have a sense of disgust with myself	(Hatred) I stop caring about myself
I feel beaten down by own self-critical thoughts	I am gentle and supportive with myself	(Hatred) I call myself names	(Hatred) I call myself names
I can't accept failures and setbacks without feeling inadequate	I do not like being me	(Hatred) I do not like being me	(Hatred) I do not like being me
		(Inadequacy) There is a part of me that feels I am not good enough	(Inadequacy) I feel beaten down by own self-critical thoughts
		(Inadequacy) I feel beaten down by own self-critical thoughts	(Inadequacy) I remember and dwell on my failings
		(Inadequacy) I can't accept failures and setbacks without feeling inadequate	(Inadequacy) I can't accept failures and setbacks without feeling inadequate
		(Self-kindness) I find it easy to forgive myself	(Self-kindness) I find it easy to forgive myself
		(Self-kindness) I still like being me	(Self-kindness) I still like being me
		(Self-kindness) I can feel lovable and acceptable	(Self-kindness) I can feel lovable and acceptable
		(Self-kindness) I am gentle and supportive with myself	(Self-kindness) I am gentle and supportive with myself

methods such as the scree-test, parallel analysis, proportion of variance accounted for by the single factor, loading size and loading range on the single EFA factor or results from a CFA model can provide reliable information. Additionally, this method does not handle ordinal or categorical data. Future developments could overcome this limitation.

In general, new machine learning techniques are currently being employed as novel ways to perform scale validation, such as for factor extraction (Goretzko, 2022; Goretzko & Bühner, 2020) or exploring alternative factor structures (Camilleri et al., 2021). Scale abbreviation represents an important sub-field in which new algorithmic developments could open up new ways to get reliable, valid, short scales and therefore ease the burden on study participants.

COMPLIANCE WITH ETHICAL STANDARDS

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee at the University of Padua and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all participants included in the study.

ACKNOWLEDGEMENT

Open access publishing facilitated by Consiglio Nazionale delle Ricerche, as part of the Wiley - CRUI-CARE agreement.

Manuscript received August 2023
Revised manuscript accepted September 2024

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1. Code for the full analysis.

REFERENCES

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, *24*(4), 718–738. https://doi.org/10.1177/1094428120947794

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322.

Brunell, A. B., Davis, M. S., Schley, D. R., Eng, A. L., van Dulmen, M. H., Wester, K. L., & Flannery, D. J. (2013). A new measure of interpersonal exploitativeness. *Frontiers in Psychology*, *4*, 299. https://doi.org/10.3389/fpsyg.2013.00299

Camilleri, J. A., Eickhoff, S. B., Weis, S., Chen, J., Amunts, J., Sotiras, A., & Genon, S. (2021). A machine learning approach for the factorization of psychometric data with application to the Delis Kaplan executive function system. *Scientific Reports*, *11*(1), 16896. https://doi.org/10.1038/s41598-021-96342-3

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151. https://doi.org/10.1177/1073191120957102

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Gilbert, P., Clarke, M., Hempel, S., Miles, J. N., & Irons, C. (2004). Criticizing and reassuring oneself: An exploration of forms, styles and reasons in female students. *British Journal of Clinical Psychology*, *43*(1), 31–50. https://doi.org/10.1348/014466504772812959

Gilbert, P., McEwan, K., Matos, M., & Rivis, A. (2011). Fears of compassion: Development of three self-report measures. *Psychology and Psychotherapy: Theory, Research and Practice*, *84*(3), 239–255. https://doi.org/10.1348/147608310X526511

Gilbert, P., McEwan, K., Mitra, R., Richter, A., Franks, L., Mills, A., Bellew, R., & Gale, C. (2009). An exploration of different types of positive affect in students and patients with a bipolar disorder. *Clinical Neuropsychiatry*, *6*(4), 135–143.

Goretzko, D. (2022). Factor retention in exploratory factor analysis with missing data. *Educational and Psychological Measurement*, *82*(3), 444–464. https://doi.org/10.1177/00131644211022031

Goretzko, D., & Bühner, M. (2020). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis. *Psychological Methods*, *25*(6), 776–786. https://doi.org/10.1037/met0000262

Iglesias, C., & Torgerson, D. (2000). Does length of questionnaire matter? A randomised trial of response rates to a mailed questionnaire. *Journal of Health Services Research & Policy*, *5*(4), 219–221. https://doi.org/10.1177/135581960000500406

Kleka, P., & Soroko, E. (2018). How to avoid the sins of questionnaires abridgement. *Survey Research Methods*, *12*(2), 147–160.

Kruyen, P. M., Emons, W. H., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, *13*(3), 223–248. https://doi.org/10.1080/15305058.2012.703734

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172–175. https://doi.org/10.1037//0033-2909.111.1.172

Neff, K. D. (2003). The development and validation of a scale to measure self-compassion. *Self and Identity*, *2*(3), 223–250. https://doi.org/10.1080/15298860309027

Neff, K. D., & McGehee, P. (2010). Self-compassion and psychological resilience among adolescents and young adults. *Self and Identity*, *9*(3), 225–240. https://doi.org/10.1080/15298860902979307

Nunnally, J. C. (1978). An overview of psychological measurement. In B. B. Wolman (Ed.), *Clinical diagnosis of mental disorders*. Springer. https://doi.org/10.1007/978-1-4684-2490-4_4

Raykov, T., & Pohl, S. (2013). Essential unidimensionality examination for multicomponent scales: An interrelationship decomposition approach. *Educational and Psychological Measurement*, *73*(4), 581–600. https://doi.org/10.1177/0013164412470451

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Richter, A., Gilbert, P., & McEwan, K. (2009). Development of an early memories of warmth and safeness scale and its relationship to psychopathology. *Psychology and Psychotherapy: Theory, Research and Practice*, *82*(2), 171–184. https://doi.org/10.1348/147608308X395213

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Sahdra, B. K., Ciarrochi, J., Parker, P., & Scrucca, L. (2016). Using genetic algorithms in a large nationally representative American sample to abbreviate the Multidimensional Experiential Avoidance Questionnaire. *Frontiers in Psychology*, *7*, 167967. https://doi.org/10.3389/fpsyg.2016.00189

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295.

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

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