information

Article

Scale Abbreviation with Recursive Feature Elimination and Genetic Algorithms: An Illustration with the Test Emotions Questionnaire

Sevilay Kilmen [1,*] and Okan Bulut [2]

[1] Faculty of Nursing, University of Alberta, Edmonton, AB T6G 1C9, Canada
[2] Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB T6G 2G5, Canada
* Correspondence: kilmen@ualberta.ca

Abstract: Psychological scales play a key role in the assessment, screening, and diagnosis of latent variables, such as emotions, mental health, and well-being. In practice, researchers need shorter scales of psychological traits to save administration time and cost. Thus, a variety of optimization algorithms have been proposed to abbreviate lengthy psychological scales into shorter instruments efficiently. The main goal of this application is to form an abbreviated scale with fewer items while maintaining reliability, relationships among the subscales, and model fit for the full scale. In this study, we use an optimization algorithm (genetic algorithm) and a feature selection algorithm (recursive feature elimination) to abbreviate a psychological scale automatically. Although both algorithms search for an optimal subset of features within a large pool of features, the search mechanism underlying each algorithm is quite different. The genetic algorithm employs a systematic but computationally-expensive sampling process to find the optimal features, whereas recursive feature elimination removes the least important features iteratively until a desired number of features are retained. In this study, we use a 77-item measure of test emotions (Test Emotions Questionnaire) to demonstrate how these algorithms can be used for scale abbreviation. We generate a 40-item short form using each algorithm and compare the quality of the selected items against the full-length scale. The results indicate that both methods can provide researchers and practitioners with a systematic procedure for creating psychometrically sound, shorter versions of lengthy psychological instruments.

Keywords: scale abbreviation; recursive feature elimination; genetic algorithms; test emotions

check for updates

Citation: Kilmen, S.; Bulut, O. Scale Abbreviation with Recursive Feature Elimination and Genetic Algorithms: An Illustration with the Test Emotions Questionnaire. *Information* **2023**, *14*, 63. https://doi.org/10.3390/info14020063

Academic Editor: Benedicenti Luigi

Received: 4 December 2022
Revised: 10 January 2023
Accepted: 17 January 2023
Published: 21 January 2023

1. Introduction

Researchers build psychological scales to measure latent variables or traits, such as intelligence, emotions, and attitudes. To cover important observable indicators of the target construct, researchers often include numerous items (i.e., questions) in their scales, leading to lengthy instruments that are time-consuming for participants. In practice, as the length of the instrument increases, participants may want to spend less time answering each item or show careless responding, deteriorating the quality of response data that fail to reflect participants' actual levels of the constructs being measured [1,2]. This situation, defined as "survey fatigue" in literature [3,4], has motivated researchers to design shortened or abbreviated forms of lengthy scales that can reduce administration time while increasing response rate [5]. Abbreviated scales allow researchers to measure participants' latent traits in much less time than required to administer the original one, without sacrificing the psychometric quality of the full-length scale.

Despite the potential advantages of using an abbreviated scale in practice, widespread use of scale abbreviation has been impeded by the laborious selection procedure researchers need to perform. Most approaches involve a tedious process where the researcher needs to

manually identify the best items to retain using the same type of conceptual and psychometric considerations employed in the development of the full-length scale [6]. For example, the researcher may perform an exhaustive search through all possible permutations of the existing items to find the best set of items based on psychometric indices (e.g., internal consistency). However, this method would require a significant amount of time and effort due to a combinatorial explosion. For example, selecting 30 items from a pool of 100 items would be:

$$^{100}C_{30} = \binom{100}{30} = \frac{100!}{30!(100-30)!}, \quad (1)$$

indicating that more than 29 septillion combinations would have to be tried, which is practically impossible. Therefore, instead of manually trying each combination, optimization methods can be implemented to find the best solution more efficiently.

Recently, researchers have adopted various optimization methods, such as genetic algorithms and ant colony optimization to solve the scale abbreviation problem, e.g., [6,7]. These algorithms are metaheuristic approaches that can be used to find approximate solutions to the problem of selecting the best items from a large pool of items. In this study, we propose "feature selection" as an alternative approach for finding the best items automatically. Our approach aims to select the best subset of features (i.e., items) that can predict respondents' scores from the full-length scale, without incurring much loss of information. Using recursive feature elimination (RFE), we fit a predictive model by using all items in the pool to predict the total scale score and recursively remove the items based on their importance until the specified number of items is retained. Using real data from a psychological scale, we first demonstrate how RFE can be utilized as a scale abbreviation method and then compare its performance to that of genetic algorithms.

2. Conceptual Framework

2.1. Recursive Feature Elimination (RFE)

Feature selection plays a vital role in machine learning and data mining tasks. Researchers employ feature selection algorithms to find a subset of features with the minimum possible generalization error or to select the smallest possible subset with a given discrimination capability [8–10]. RFE is a widely used algorithm for selecting features that are most relevant in predicting the target variable in a predictive model (either regression or classification) [11]. This method implements a backward selection process to find the optimal combination of features by eliminating non-predictive or redundant features. First, it builds a predictive model based on all features and then calculates the importance of each feature. Second, it rank-orders the features and identifies relatively less important or redundant features. Finally, it removes the features with the least importance recursively based on model evaluation metrics (e.g., accuracy, Kappa, and root mean squared error) until a desired number of features remains in the model.

RFE is a wrapper-type feature selection algorithm that can utilize a variety of machine learning algorithms to select the best features. To date, researchers have used RFE with several machine learning algorithms, such as Least Absolute Shrinkage and Selection Operator (LASSO) and ridge regression, support vector machine, and Random Forest [12,13]. However, recent research suggests that RFE combined with the Random Forest algorithm provides more stable results with improved accuracy [9,14,15]. First, Random Forest is used to build a predictive model and calculate the importance of the features using a particular method, such as classification accuracy of "out-of-bag" data. Second, the features are ranked based on their importance and the least important feature(s) are eliminated from the list of available features. Third, the remaining features are used to make a prediction using Random Forest [16]. This recursive process continues until RFE identifies the most important features to be retained in the model [8,12,16]. In the context of scale abbreviation, "features" refer to the individual items in the scale and the target variable is the total (raw) scores computed based on the full set of items. Using Random Forest for the prediction of

the total scores, RFE can iteratively search for the "most important" items to be retained in the abbreviated form [17].

2.2. Genetic Algorithms

Genetic algorithms are a metaheuristic optimization approach inspired by the process of natural selection that belongs to evolutionary algorithms [18]. This approach is based on the concept of "survival of the fittest" in Darwin's theory of evolution [19]. Genetic algorithms mimic biological processes (e.g., mutation, crossover, and selection) to generate high-quality solutions to optimization and search problems [18]. This method essentially aims to decrease redundancy within a situation by reducing the selection of the items for the substrate that captures the traits of interest [20,21]. Therefore, it can be used as an optimization algorithm to solve various selection problems, such as scale abbreviation. In the scale abbreviation context, genetic algorithms consider each item as a single gene on a "chromosome" containing all items in the scale. On these chromosomes, the retained items are coded as 1 while the remaining items are coded as 0. These codes are called "genes" (see Figure 1). For example, if the target scale consists of 10 items and 4 items are to be selected, then one of the possible lineups would be "1001100010". This particular lineup shows that the first, fourth, fifth, and ninth items will be retained in the abbreviated form, whereas the rest will be eliminated [20].

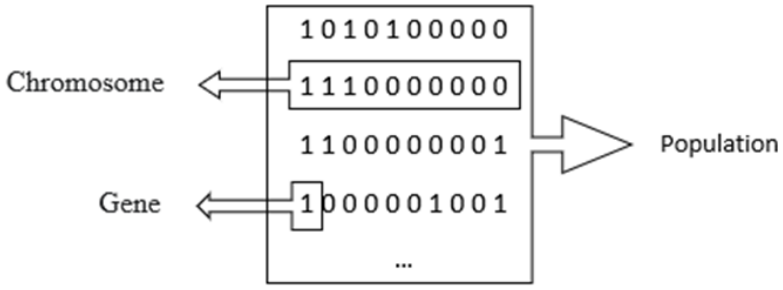

Figure 1. Gene, chromosome, and population in genetic algorithms.

The chromosome community is known as a population in genetic algorithms. Genetic algorithms select a random initial population represented in different points in the search space. This population is called the first generation. Then, fit chromosomes are transferred to the next population (i.e., next generations). The degree of fitness of the chromosomes is evaluated using the fitness function. The principle of survival in genetic algorithms depends on the results of the fitness function [22]. The fitness function for a psychological scale measuring a single latent trait can be written as follows [6,23]:

$$Cost = Ik + (1 - R^2), \quad (2)$$

where I is a fixed item cost determined by the researcher, k is the number of items to be retained in the abbreviated scale, and R^2 is the amount of total explained variance explained by a linear combination of item scores. Equation (2) shows that the genetic algorithms aim to balance the cost of each additional item while maximizing the amount of explained variance in the scale. The genetic algorithm's selection-reproduction scheme is implemented to identify chromosomes based on the fitness function shown in Equation (2). The primary genetic operators applied here are crossover and mutation. In a crossover process (see Figure 2), design characteristics between any paired individuals are exchanged to form two new child chromosomes. In this process, two chromosomes (parents) are chosen by giving a higher probability of selection to chromosomes with a small fitness value [24]. Then, two new offspring chromosomes are produced by using the parent chromosomes.

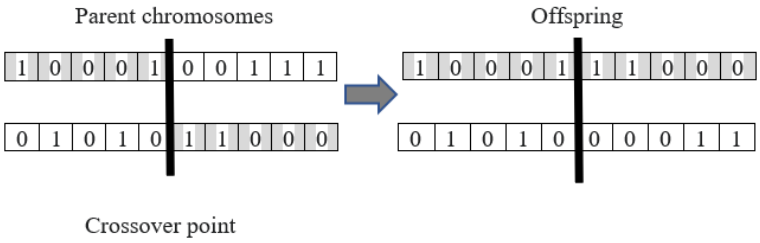

Figure 2. Crossover process in genetic algorithms.

Following crossover, the mutation (see Figure 3) is applied to the genes of child chromosomes by changing a gene of 1 to 0 or 0 to 1 for a random investigation of the design space [24]. The selection process, crossover, and mutation continue until a termination condition is satisfied [22]. Finally, at the end of the iterative process, the genetic algorithms determine the items that would compose the optimal static short form.

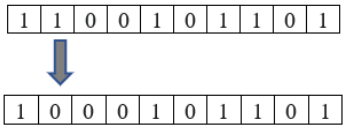

Figure 3. Mutation.

2.3. Current Study

In this study, we use the Test Emotions Questionnaire (TEQ), which is a subscale of the Achievement Emotions Questionnaire [25], to illustrate how RFE and genetic algorithms can be used for abbreviating psychological scales. Test emotions refer to a set of emotions that individuals may feel with regard to taking tests and exams [26]. Following several studies focusing on test emotions, Pekrun and his colleagues [26] identified eight primary test emotions that individuals are likely to experience when taking a test. These emotions are anger, anxiety, enjoyment, hopelessness, hope, pride, relief, and shame. Pekrun and his colleagues [26] designed the TEQ to perform a comprehensive assessment of these test-related emotions. The TEQ consists of 77 Likert-scale items from 8 subscales focusing on distinct emotions related to the test-taking process. The TEQ has been widely used in empirical studies focusing on test-related emotions and adapted to different cultures [27–30].

As the TEQ consists of many items (i.e., 77 items), its administration time ranges from 50 to 70 minutes. Since the TEQ is not suitable for individuals who may not be motivated to complete such a lengthy instrument, Lichtenfeld and her colleagues [31] designed an abbreviated version of the TEQ for both German and American elementary student samples. Focusing on the primary test-related emotions (i.e., enjoyment and anxiety), they measured enjoyment with three items and anxiety with five items. Similarly, Peixoto and colleagues [30] also developed a short form of the TEQ with 24 items for a Portuguese secondary student sample. They focused on six emotions (anger, anxiety, enjoyment, hopelessness, pride, and relief) out of the eight test emotions and used four items to measure each emotion. Since neither of these abbreviated forms of the TEQ were intended to be used for undergraduate students, Bieleke and colleagues [32] developed a 32-item form of TEQ for a Canadian undergraduate student sample. However, there is no evidence about cultural invariance of the abbreviated TEQ beyond the Canadian undergraduate student population. In this study, we use response data from a sample of Turkish undergraduate students to build an abbreviated form of the TEQ using RFE and genetic algorithms. Our goal is to illustrate how these two methods can be used for scale abbreviation, while comparing the performance of these methods in building an accurate, abbreviated forms of the TEQ.

3. Materials and Methods

3.1. Participants

The sample of this study consisted of 559 undergraduate students from a university located in the northwest of the Black Sea region of Turkey who consented to participate in the study voluntarily. The students ranging from 18 to 28 years old ($M = 20.7$ $SD = 1.65$) completed the Turkish version of the TEQ with 77 items. The administration time of the TEQ was approximately 60 minutes on average. Less than 5% of the students ($n = 16$) skipped some items in the scale, whereas the remaining students completed all the items in the TEQ. The students skipping the items were excluded from the subsequent analyses, and, thus, the final dataset consisted of 543 students. All study procedures were approved by the research ethics board of the first author's academic institution.

3.2. Instrument

The TEQ is a self-report scale developed by Pekrun and colleagues [25,26]. It was embedded within the Achievement Emotions Questionnaire [33] with three subscales (i.e., class-related emotions, learning-related emotions, and test-related emotions). The TEQ measures test-related emotions with 77 items (see Table A1 in Appendix A for the TEQ subscales and their items). Participants in the current study responded to the items using a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). It consists of eight dimensions: anger (10 items, e.g., I am fairly annoyed after taking a test), anxiety (12 items, e.g., Before taking a test, I feel nervous and uneasy), enjoyment (10 items, e.g., I am happy that I can cope with the test during taking a test), hopelessness (11 items, e.g., My hopelessness robs me of all my energy before taking an exam), hope (8 items, e.g., I think about my exam optimistically before taking a test), pride (10 items, e.g., I am proud of how well I mastered a test after taking it), relief (6 items, e.g., I think that I finally can breathe easy again after taking a test), and shame (10 items, e.g., I get so embarrassed I want to run and hide during taking an exam). Can and colleagues [27] adapted the TEQ into the Turkish language. The results obtained from Can and colleagues' study showed that the TEQ is a reliable and valid scale to measure test emotions within the Turkish culture.

3.3. Data Analysis

Our preliminary analysis indicated that less than 3% of the students ($n = 16$) did not answer some of the items in the TEQ, and, thus, these students were removed from the final dataset. To obtain abbreviated forms of the TEQ (i.e., 5 items for each dimension of the TEQ; a 40-item scale in total) based on the genetic algorithms and RFE approaches, we used the GAabbreviate [34], randomForest [35], and caret [36] packages in R [37]. We evaluated the quality of the abbreviated forms using four criteria. First, we used the coefficient alpha (α) [38] to examine and compare the internal consistency (i.e., reliability) of the full-length and abbreviated forms of the TEQ. Second, we calculated the relationships between the total raw scores obtained from the full-length and abbreviated forms of the TEQ using the Pearson correlation coefficient. To avoid spurious inflation in the correlation due to shared error variance between the full-length and abbreviated forms, we performed a correction procedure proposed by Levy [39] based on the reliability of the full-length and abbreviated forms. Third, we reviewed the correlations among the eight subscales of the TEQ for the full-length and abbreviated forms. Finally, we examined the model fit of the original and short forms of the TEQ by fitting a one-factor confirmatory factor analysis (CFA) model for each TEQ sub-scale separately. The CFA models were estimated using the unweighted least squares method available in the lavaan package [40]. The model data fit was evaluated using the Comparative Fit Index (Good fit: CFI $\geq$ 0.90), Tucker–Lewis Index (Good fit: TLI $\geq$ 0.90), and Standardize Root Mean Square Residual (Good fit: SRMR $\leq$ 0.05) [41–44].

4. Results

The results revealed that except for the relief subscale, RFE and genetic algorithms selected different items for the 5-item abbreviated forms of the TEQ subscales (see Table 1).

The following sections summarize the results based on the four evaluation criteria explained above: (1) reliability, (2) relationship with the full-length scale, (3) relationships among the subscales, and (4) model fit indices.

Table 1. Reliability and correlation results for the full-length and abbreviated subscales of the TEQ.

Test Emotions	Abbreviation Approach	Selected Items	α	r
Anger	RFE	1, 2, 6, 9, 10	0.728	0.719
	GA	1, 4, 6, 7, 10	0.646	0.737
	Full		0.783	-
Anxiety	RFE	3, 6, 9, 11, 12	0.792	0.832
	GA	1, 2, 5, 9, 12	0.781	0.847
	Full		0.841	-
Enjoyment	RFE	1, 6, 8, 9, 10	0.647	0.716
	GA	5, 6, 7, 8, 9	0.729	0.757
	Full		0.820	-
Hope	RFE	1, 3, 4, 6, 8	0.707	0.765
	GA	2, 3, 5, 6, 7	0.728	0.786
	Full		0.802	-
Hopelessness	RFE	2, 6, 8, 9, 10	0.719	0.789
	GA	1, 3, 5, 9, 11	0.804	0.848
	Full		0.870	-
Pride	RFE	2, 4, 5, 8, 9	0.819	0.859
	GA	2, 4, 7, 8, 9	0.783	0.869
	Full		0.853	-
Relief	RFE	1, 2, 4, 5, 6	0.658	0.683
	GA	1, 2, 4, 5, 6	0.658	0.683
	Full		0.702	-
Shame	RFE	1, 3, 4, 7, 9	0.777	0.813
	GA	4, 6, 7, 8, 9	0.770	0.834
	Full		0.871	-

Note: RFE = Recursive feature elimination. GA = Genetic algorithms. Full = Full-length scale. α refers to the coefficient alpha. r is the corrected correlation between the long and short forms.

4.1. Reliability

Table 1 shows that the α values obtained from the full-length subscales are higher than those for the abbreviated subscales. This finding is not necessarily surprising because, as the number of items decreases, α also tends to decrease due to the reduced variation in the abbreviated scale [45]. To evaluate the reliability level of each subscale, we followed George and Mallery's criteria (i.e., $\alpha > 0.90$ excellent, >0.80 good, >0.70 acceptable, >0.60 questionable, >0.50 poor, and <0.50 unacceptable; [46]). Based on these criteria, the subscales of enjoyment (abbreviated by RFE), anger (abbreviated by genetic algorithms), and relief (abbreviated by both methods identically) failed to indicate an "acceptable" level of reliability. RFE outperformed genetic algorithms in four of the TEQ subscales, whereas genetic algorithms yielded higher α values in three TEQ subscales.

There are two noteworthy findings. First, the reliability difference between the abbreviated forms seems to be much larger in the three subscales where genetic algorithms performed better (e.g., for hopelessness, genetic algorithms: $\alpha = 0.804$, RFE: $\alpha = 0.719$). In contrast, the two methods generally yielded similar α values in the subscales for which RFE seemed superior. Second, RFE and genetic algorithms selected the same items for

the relief subscale by eliminating the same item (i.e., item 3) from the full-length subscale. Abbreviating this subscale seemed detrimental to reliability as it lowered α.

4.2. Relationship with the Full-Length Scale

The second evaluation criterion was the relationship between the abbreviated and full-length subscales of the TEQ. If the scale abbreviation process works as expected, then total scores from the abbreviated subscales should have a high correlation with those from the full-length subscales. The results in Table 1 show that the abbreviated subscales obtained from both methods indicated strong correlations (corrected for spurious inflation) with the full-length subscales, ranging from $r = 0.683$ to $r = 0.869$. Except for the relief subscale for which both methods selected the same items, the abbreviated subscales from genetic algorithms indicated slightly higher correlations with the full-length subscales than those from RFE. Overall the results showed that if the participants completed the abbreviated subscales (40 items in total) rather than the full-length scales (77 items in total), their relative positions in the sample in terms of their test emotions would not change significantly.

4.3. Correlations among the Subscales

The third evaluation criterion was the relationships (i.e., correlations) among the eight subscales of the TEQ. Although each subscale of the TEQ measures a different aspect of test emotions, the results of these subscales are likely to correlate, given the similarities between the test emotions. If the scale abbreviation process works properly, the correlations among the abbreviated subscales should be similar to those from the full-length subscales. Figure 4 shows the correlation matrix plot for the full-length subscales (left), as well as those for genetic algorithms (middle) and RFE (right). The results show that the abbreviated subscales obtained from both genetic algorithms and RFE maintained the relationships among the subscales accurately. However, for some pairwise relationships (e.g., enjoyment-hopelessness, pride-hopelessness, and anger-enjoyment), the RFE-based subscales produced weaker correlations, suggesting that the abbreviated subscales from RFE could not maintain the existing relationships in the full-length scale. Similarly, genetic algorithms also failed to maintain some pairwise relationships in the full-length scale (e.g., pride-anger and pride-anxiety).

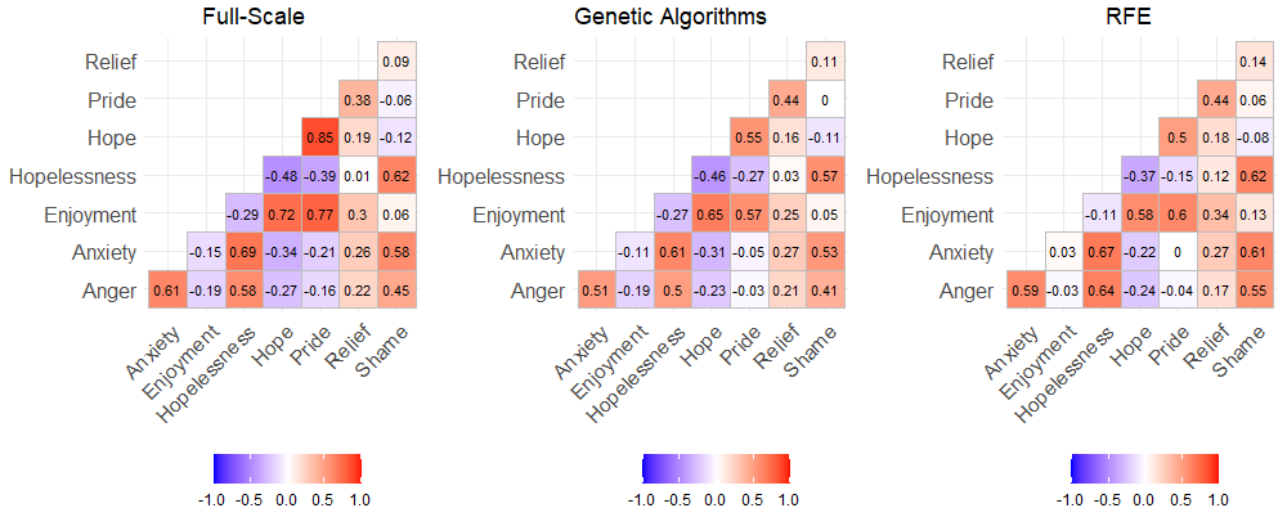

Figure 4. Correlations among the subscales of the TEQ.

4.4. Model Data Fit

Our last evaluation criterion was the model-data fit for the abbreviated subscales of the TEQ. Each subscale of the TEQ measures a unidimensional (i.e., one-factor) construct related to test emotions (e.g., anger, anxiety, and enjoyment). After the subscales are

abbreviated, they are expected to maintain the same unidimensional structure. That is, a one-factor CFA model should indicate adequate fit for the abbreviated subscales. Table 2 shows the model fit indices obtained from the full-length and abbreviated subscales of the TEQ. All the abbreviated subscales obtained from genetic algorithms and RFE yielded reasonable CFI (0.939 or higher), TLI (0.844 or higher), and SRMR (0.073 or smaller) values. For some subscales (e.g., anger and anxiety), the abbreviated forms yielded even better fit indices than the full-length subscales. Generally, RFE produced better results than genetic algorithms across the three model fit indices.

Table 2. CFA results for the full-length and abbreviated subscales of the TEQ.

Test Emotions	Approach	CFI	TLI	SRMR
Anger	RFE	0.991	0.983	0.035
	GA	0.939	0.877	0.059
	Full	0.929	0.908	0.074
Anxiety	RFE	0.995	0.998	0.034
	GA	0.986	0.972	0.049
	Full	0.974	0.967	0.063
Enjoyment	RFE	0.992	0.844	0.073
	GA	0.992	0.984	0.032
	Full	0.964	0.954	0.061
Hope	RFE	0.962	0.924	0.059
	GA	1	1	0.015
	Full	0.986	0.981	0.051
Hopelessness	RFE	1	1	0.017
	GA	0.992	0.984	0.040
	Full	0.991	0.989	0.046
Pride	RFE	0.999	0.999	0.023
	GA	0.997	0.993	0.029
	Full	0.984	0.959	0.054
Relief	RFE	1	1	0.020
	GA	1	1	0.020
	Full	0.994	0.991	0.032
Shame	RFE	0.996	0.992	0.023
	GA	0.996	0.993	0.028
	Full	0.991	0.988	0.048

Note: RFE = Recursive feature elimination, GA = Genetic algorithms, Full = Full-length scale.

5. Discussion

In psychological, educational, and sociological research, researchers often use lengthy instruments to assess psychological constructs, such as personality types, emotions, and mental disorders, with a high degree of accuracy. Since this approach requires participants to invest a significant amount of time in completing the instrument, various aberrant response behaviors (e.g., careless and insufficient responding) may be observed. The presence of aberrant responses may contaminate the results and pose a major threat to the validity of inferences drawn from the instrument [47]. Therefore, there is a growing demand for short or abbreviated scales that can be used for research purposes and in making individual-level decisions, especially in clinical settings (e.g., screening for mental disorders) and personnel selection [48]. Using an abbreviated scale can help researchers limit the administration time and reduce the possibility of aberrant response behaviors while improving the validity of inferences being made from the scale.

To date, researchers have proposed various metaheuristic approaches for abbreviating length scales efficiently. In this study, we used one of these metaheuristic approaches (genetic algorithms) and a feature selection algorithm (RFE) to abbreviate a 77-item scale into a 40-item abbreviated scale. To our knowledge, no study has utilized recursive

feature elimination to abbreviate psychological scales. In this study, we performed scale abbreviation using genetic algorithms and RFE and then compared their performance based on several criteria, including scale reliability, relationship with the full-length scale, relationships among the subscales, and model fit indices. Our results showed that both genetic algorithms and RFE returned abbreviated subscales of the TEQ that had lower coefficient alpha values (i.e., lower reliability) than the full-length scale. Considering that the coefficient alpha tends to underestimate reliability for short and/or heterogeneous measures [49], this finding is not necessarily surprising. Across the eight subscales of the TEQ, genetic algorithms yielded coefficient alpha values that are either higher than or very similar to those obtained from RFE.

Despite the reduced reliability values, the abbreviated subscales obtained from both methods indicated strong correlations with the full-length subscales. This finding has two implications. First, the relative positions of the participants in the data would not change significantly if the abbreviated subscales, rather than the full-length instrument, were used. Second, the results of the abbreviated subscales could predict the full-length instrument with a high accuracy. In terms of maintaining the relationships among the subscales in the full-length instrument, genetic algorithms generally outperformed RFE. However, both methods yielded weaker correlations among subscales, compared to those from the full-length instrument. This is very likely to be a consequence of reduced variation (i.e., heterogeneity) in the abbreviated scales. Lastly, the abbreviated forms yielded better model fit indices than the subscales from the full-length version of the TEQ. This result suggests that as genetic algorithms and RFE abbreviated the subscales of the TEQ, the items not contributing to the model fit were eliminated. In addition, scale abbreviation with both genetic algorithms and RFE was computationally simple as both methods returned the abbreviated subscales quickly.

Our findings showed that genetic algorithms and RFE can produce abbreviated scales that can (a) produce reliable results, (b) predict the results of the full-length instrument accurately, (c) maintain the relationships among the subscales, and (d) select the items consistently within a unidimensional factor structure. The current study also provided additional evidence that metaheuristic methods, such as genetic algorithms, are robust approaches for abbreviating lengthy psychological instruments into a shorter one, yielding psychometrically-sound instruments for researchers and practitioners, e.g., [6,22,23]. Additionally, feature selection algorithms, such as RFE, seem to be a promising approach for abbreviating lengthy scales efficiently. Overall, both genetic algorithms and RFE can be used for selecting the best items automatically with minimal effort. Abbreviated scales produced by genetic algorithms and RFE can help reduce the administration time and avoid unintended consequences such as careless responding caused by cognitive fatigue related to the length of the instrument.

6. Limitations and Recommendations

This study has several limitations. First, we performed scale abbreviation with RFE based on the Random Forest algorithm and its variable importance mechanism. However, there are other machine learning algorithms (e.g., Lasso regression, decision trees, and neural networks) and metrics for evaluating feature importance (e.g., ROC curves). Future studies can evaluate the performance of RFE with different predictive algorithms and feature importance metrics. Second, the RFE method requires building a predictive model using a target variable and then determines feature importance based on the power of features in predicting the target variable accurately. In this study, we used raw scores calculated based on the full set of items as the target variable for illustrative purposes. Given the ordinal nature of the Likert response scale in the TEQ, calculating the subscale scores based on the item response theory framework could yield more stable scores to be used in the predictive model stage of RFE. Third, we used psychometric and statistical metrics (e.g., reliability, correlations with the full-length scale, and model fit) to compare the performance of genetic algorithms and RFE in abbreviating the subscales of the TEQ.

However, other evaluation criteria (e.g., measurement invariance of the abbreviated subscales across demographic variables, such as gender) can also be utilized for making a more comprehensive comparison between the two methods in future research. Lastly, we found that some correlations between the abbreviated and full-length subscales exceeded the maximum possible value (i.e., the square root of the product of the reliability levels of the tests) [50]. Nimon et al. [51] explained that this issue occurs in the presence of nuisance correlations between the error scores from the two tests. Therefore, it is likely that the reported correlations between the abbreviated and full-length subscales might be somewhat inflated, despite using Levy [39]'s correction for eliminating spurious correlations. Further research should focus on developing more robust methods for finding true correlations between the abbreviated and full-length forms of psychological instruments.

Author Contributions: Conceptualization, S.K. and O.B.; methodology, S.K. and O.B.; software, S.K. and O.B.; validation, S.K. and O.B.; formal analysis, S.K.; investigation, S.K. and O.B.; data curation, S.K.; writing—original draft preparation, S.K. and O.B.; writing—review and editing, S.K. and O.B.; visualization, O.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects in this study.

Data Availability Statement: Data and R codes are freely available at https://osf.io/jzdhn/ (accessed on 10 January 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CFA	Confirmatory factor analysis
CFI	Comparative Fit Index
RFE	Recursive feature elimination
SRMR	Standardize Root Mean Square Residual
TEQ	Test Emotions Questionnaire
TLI	Tucker–Lewis Index

Appendix A

Table A1. Subscales and items of the full-length TEQ.

Item	Subscale	Content
1	Anger	I get angry over time pressures which don't leave enough time to prepare. (b)
2	Anger	I get angry about the amount of material I need to know. (b)
3	Anger	I get angry. (d)
4	Anger	I think the questions are unfair. (d)
5	Anger	I get angry about the teacher's grading standards. (a)
6	Anger	I am fairly annoyed. (a)
7	Anger	I wish I could tell the teacher off. (a)
8	Anger	I wish I could freely express my anger. (a)
9	Anger	My anger makes the blood rush to my head. (a)
10	Anger	I get so angry, I start feeling hot and flushed. (a)
1	Anxiety	I worry whether I have studied enough. (b)
2	Anxiety	I feel sick to my stomach. (b)
3	Anxiety	Before the exam, I feel nervous and uneasy. (b)
4	Anxiety	I get so nervous I wish I could just skip the exam. (b)
5	Anxiety	I worry whether the test will be too difficult. (b)
6	Anxiety	I worry whether I will pass the exam. (d)

Table A1. *Cont.*

Item	Subscale	Content
7	Anxiety	At the beginning of the test, my heart starts pounding. (d)
8	Anxiety	I am very nervous. (d)
9	Anxiety	My hands get shaky. (d)
10	Anxiety	I get so nervous I can't wait for the exam to be over. (d)
11	Anxiety	I feel panicky when writing the exam. (d)
12	Anxiety	I am so anxious that I'd rather be anywhere else. (d)
1	Enjoyment	I look forward to the exam. (b)
2	Enjoyment	Because I enjoy preparing for the test, I'm motivated to do more than is necessary. (b)
3	Enjoyment	Before taking the exam, I sense a feeling of eagerness. (b)
4	Enjoyment	I look forward to demonstrating my knowledge. (b)
5	Enjoyment	Because I look forward to being successful, I study hard. (b)
6	Enjoyment	I enjoy taking the exam. (d)
7	Enjoyment	I am happy that I can cope with the test. (d)
8	Enjoyment	For me, the test is a challenge that is enjoyable. (d)
9	Enjoyment	My heart beats faster with joy. (a)
10	Enjoyment	I glow all over. (a)
1	Hope	I start studying for the exam with great hope and anticipation. (b)
2	Hope	I am optimistic that everything will work out fine. (b)
3	Hope	I have great hope that my abilities will be sufficient. (b)
4	Hope	I'm quite confident that my preparation is sufficient. (b)
5	Hope	I think about my exam optimistically. (b)
6	Hope	My confidence motivates me to prepare well. (b)
7	Hope	Hoping for success, I'm motivated to invest a lot of effort. (d)
8	Hope	I am very confident. (d)
1	Hopelessness	My hopelessness robs me of all my energy. (b)
2	Hopelessness	I have lost all hope that I have the ability to do well on the exam. (b)
3	Hopelessness	I feel so resigned about the exam that I can't start doing anything. (b)
4	Hopelessness	I'd rather not write the test because I have lost all hope. (b)
5	Hopelessness	I get depressed because I feel I don't have much hope for the exam. (b)
6	Hopelessness	I start to think that no matter how hard I try, I won't succeed on the test. (d)
7	Hopelessness	I feel like giving up (d)
8	Hopelessness	I start to realize that the questions are much too difficult for me. (d)
9	Hopelessness	I feel so resigned that I have no energy. (d)
10	Hopelessness	I have given up believing that I can answer the questions correctly. (d)
11	Hopelessness	I feel hopeless. (d)
1	Pride	I'm so proud of my preparation that I want to start the exam now. (b)
2	Pride	I think that I can be proud of my knowledge. (d)
3	Pride	Pride in my knowledge fuels my efforts in doing the test. (d)
4	Pride	When I get the test results back, my heart beats with pride. (a)
5	Pride	I'm proud of how well I mastered the exam. (a)
6	Pride	To think about my success makes me feel proud. (a)
7	Pride	After the exam, I feel ten feet taller because I'm so proud. (a)
8	Pride	I am very satisfied with myself. (a)
9	Pride	I walk out of the exam with the look of a winner on my face. (a)
10	Pride	I am proud of myself. (a)
1	Relief	The tension in my stomach is dissipated. (a)
2	Relief	I finally can breathe easy again. (a)I feel freed. (a)
3	Relief	I feel very relieved. (a)
4	Relief	I feel relief. (a)
5	Relief	I can finally laugh again. (a)
6	Relief	I can finally laugh again. (a)

Table A1. *Cont.*

Item	Subscale	Content
1	Shame	I can't even think about how embarrassing it would be to fail the exam. (b)
2	Shame	I am ashamed of my poor preparation. (d)
3	Shame	I feel humiliated. (d)
4	Shame	I get so embarrassed I want to run and hide. (d)
5	Shame	Because I am ashamed, my pulse races. (d)
6	Shame	I get embarrassed because I can't answer the questions correctly. (d)
7	Shame	I feel ashamed. (a)
8	Shame	My marks embarrass me. (a)
9	Shame	When I get a bad mark, I would prefer not to face my teacher again. (a)
10	Shame	When others find out about my poor marks, I start to blush. (a)

Note: b = before taking a test/exam, d = before taking a test/exam, a = before taking a test/exam.

References

1. Backor, K.; Golde, S.; Nie, N. Estimating survey fatigue in time use study. In *Proceedings of the International Association for Time Use Research Conference*; Citeseer: Washington, DC, USA, 2007.
2. Ward, M.; Meade, A.W. Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annu. Rev. Psychol.* **2022**, *74*. [CrossRef] [PubMed]
3. Sinickas, A. Finding a cure for survey fatigue. *Strateg. Commun. Manag.* **2007**, *11*, 11.
4. Whitcomb, W.; Weitzer, W.; Porter, S. Multiple surveys of students and survey fatigue. *New Dir. Institutional Res.* **2004**, *12*, 63–73.
5. Schoeni, R.F.; Stafford, F.; McGonagle, K.A.; Andreski, P. Response rates in national panel surveys. *Ann. Am. Acad. Political Soc. Sci.* **2013**, *645*, 60–87. [CrossRef] [PubMed]
6. Yarkoni, T. The abbreviation of personality, or how to measure 200 personality scales with 200 items. *J. Res. Personal.* **2010**, *44*, 180–198. [CrossRef]
7. Leite, W.L.; Huang, I.C.; Marcoulides, G.A. Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivar. Behav. Res.* **2008**, *43*, 411–431. [CrossRef] [PubMed]
8. Zhou, Q.; Zhou, H.; Zhou, Q.; Yang, F.; Luo, L. Structure damage detection based on random forest recursive feature elimination. *Mech. Syst. Signal Process.* **2014**, *46*, 82–90. [CrossRef]
9. Granitto, P.M.; Furlanello, C.; Biasioli, F.; Gasperi, F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemom. Intell. Lab. Syst.* **2006**, *83*, 83–90. [CrossRef]
10. Lu, X.; Yang, Y.; Wu, F.; Gao, M.; Xu, Y.; Zhang, Y.; Yao, Y.; Du, X.; Li, C.; Wu, L.; et al. Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images. *Medicine* **2016**, *95*, e3973. [CrossRef]
11. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
12. Singh, S.; Agrawal, A.; Kodamana, H.; Ramteke, M. Multi-objective optimization based recursive feature elimination for process monitoring. *Neural Process. Lett.* **2021**, *53*, 1081–1099. [CrossRef]
13. Hamada, M.; Tanimu, J.J.; Hassan, M.; Kakudi, H.A.; Robert, P. Evaluation of Recursive Feature Elimination and LASSO Regularization-based optimized feature selection approaches for cervical cancer prediction. In Proceedings of the 2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), Singapore, 20–23 December 2021; pp. 333–339.
14. Chen, Q.; Meng, Z.; Liu, X.; Jin, Q.; Su, R. Decision variants for the automatic determination of optimal feature subset in RF-RFE. *Genes* **2018**, *9*, 301. [CrossRef] [PubMed]
15. Pullanagari, R.R.; Kereszturi, G.; Yule, I. Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regression. *Remote Sens.* **2018**, *10*, 1117. [CrossRef]
16. Shang, Q.; Feng, L.; Gao, S. A hybrid method for traffic incident detection using random forest-recursive feature elimination and long short-term memory network with Bayesian optimization algorithm. *IEEE Access* **2020**, *9*, 1219–1232. [CrossRef]
17. Bulut, O. How to Shorten a Measurement Instrument Automatically (Part I). 2021. Available online: https://okan.cloud/posts/2021-01-04-how-to-shorten-a-measurement-instrument-automatically-part-i/ (accessed on 10 January 2023).
18. Mitchell, M. *An Introduction to Genetic Algorithms*; MIT Press: Cambridge, MA, USA, 1998.
19. Chang, T.J.; Yang, S.C.; Chang, K.J. Portfolio optimization problems in different risk measures using genetic algorithm. *Expert Syst. Appl.* **2009**, *36*, 10529–10537. [CrossRef]
20. Eisenbarth, H.; Lilienfeld, S.O.; Yarkoni, T. Using a genetic algorithm to abbreviate the Psychopathic Personality Inventory–Revised (PPI-R). *Psychol. Assess.* **2015**, *27*, 194. [CrossRef] [PubMed]
21. Whitley, D.; Tutorial, A.G.A. Statistics and computing, 4. *Kluwer Acad. Publ.* **1994**, *4*, 65–85.
22. Rachmani, E.; Hsu, C.Y.; Nurjanah, N.; Chang, P.W.; Shidik, G.F.; Noersasongko, E.; Jumanto, J.; Fuad, A.; Ningrum, D.N.A.; Kurniadi, A.; et al. Developing an Indonesia's health literacy short-form survey questionnaire (HLS-EU-SQ10-IDN) using the feature selection and genetic algorithm. *Comput. Methods Programs Biomed.* **2019**, *182*, 105047. [CrossRef]

23. Schroeders, U.; Wilhelm, O.; Olaru, G. Meta-Heuristics in Short Scale Construction: Ant Colony Optimization and Genetic Algorithm. *PLoS ONE* **2016**, *11*, e0167110. [CrossRef]
24. Hasançebi, O.; Erbatur, F. Evaluation of crossover techniques in genetic algorithm based optimum structural design. *Comput. Struct.* **2000**, *78*, 435–448. [CrossRef]
25. Pekrun, R.; Goetz, T.; Frenzel, A.C.; Barchfeld, P.; Perry, R.P. Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemp. Educ. Psychol.* **2011**, *36*, 36–48. [CrossRef]
26. Pekrun, R.; Goetz, T.; Perry, R.P.; Kramer, K.; Hochstadt, M.; Molfenter, S. Beyond test anxiety: Development and validation of the Test Emotions Questionnaire (TEQ). *Anxiety, Stress Coping* **2004**, *17*, 287–316. [CrossRef]
27. Can, Y.; Bardakci, S.; Sarikaya, E.E. The Effect of Using Student Response System on Achievement and Achievement Emotions in An English Course. *Technol. Knowl. Learn.* **2021**, 1–37. [CrossRef]
28. Dermitzaki, I.; Bonoti, F.; Kriekouki, M. Examining test emotions in university students: Adaptation of the test emotions questionnaire in the Greek language. *Hell. J. Psychol.* **2016**, *13*, 93–115.
29. Datu, J.A.D.; Fong, R.W. Examining the association of grit with test emotions among Hong Kong Chinese primary school students. *Sch. Psychol. Int.* **2018**, *39*, 510–525. [CrossRef]
30. Peixoto, F.; Mata, L.; Monteiro, V.; Sanches, C.; Pekrun, R. The achievement emotions questionnaire: Validation for pre-adolescent students. *Eur. J. Dev. Psychol.* **2015**, *12*, 472–481. [CrossRef]
31. Lichtenfeld, S.; Pekrun, R.; Stupnisky, R.H.; Reiss, K.; Murayama, K. Measuring students' emotions in the early years: the achievement emotions questionnaire-elementary school (AEQ-ES). *Learn. Individ. Differ.* **2012**, *22*, 190–201. [CrossRef]
32. Bieleke, M.; Gogol, K.; Goetz, T.; Daniels, L.; Pekrun, R. The AEQ-S: A short version of the Achievement Emotions Questionnaire. *Contemp. Educ. Psychol.* **2021**, *65*, 101940. [CrossRef]
33. Pekrun, R.; Goetz, T.; Titz, W.; Perry, R.P. Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educ. Psychol.* **2002**, *37*, 91–105. ._4. [CrossRef]
34. Scrucca, L.; Sahdra, B.K.; Sahdra, M.B.K. Package 'GAabbreviate'. 2016. Available online: https://cran.r-project.org/web/packages/GAabbreviate/ (accessed on 10 January 2023).
35. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
36. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; Team, R.C.; et al. Package 'caret'. *R. J.* **2020**, *223*, 7.
37. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
38. Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika* **1951**, *16*, 297–334. [CrossRef]
39. Levy, P. The correction for spurious correlation in the evaluation of short-form tests. *J. Clin. Psychol.* **1967**, *23*, 84–86. .:1<84::aid-jclp2270230123>3.0.co;2-2. [CrossRef]
40. Rosseel, Y. lavaan: An R Package for Structural Equation Modeling. *J. Stat. Softw.* **2012**, *48*, 1–36. [CrossRef]
41. Baumgartner, H.; Homburg, C. Applications of structural equation modeling in marketing and consumer research: A review. *Int. J. Res. Mark.* **1996**, *13*, 139–161. [CrossRef]
42. Bentler, P.M. Multivariate analysis with latent variables: Causal modeling. *Annu. Rev. Psychol.* **1980**, *31*, 419–456. [CrossRef]
43. Bentler, P.M.; Bonett, D.G. Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bull.* **1980**, *88*, 588. [CrossRef]
44. Browne, M.W.; Cudeck, R. Alternative ways of assessing model fit. *Sociol. Methods Res.* **1992**, *21*, 230–258. [CrossRef]
45. Schmitt, N. Uses and abuses of coefficient alpha. *Psychol. Assess.* **1996**, *8*, 350. [CrossRef]
46. George, D.; Mallery, P. *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference*; Routledge: London, UK, 2019.
47. Ulitzsch, E.; Yildirim-Erbasli, S.N.; Gorgun, G.; Bulut, O. An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *Br. J. Math. Stat. Psychol.* **2022**, *75*, 668–698. .: 10.1111/bmsp.12272. [CrossRef]
48. Ziegler, M.; Kemper, C.J.; Kruyen, P. Short scales–Five misunderstandings and ways to overcome them. *J. Individ. Differ.* **2014**, *35*, 185–189. [CrossRef]
49. Osburn, H.G. Coefficient alpha and related internal consistency reliability coefficients. *Psychol. Methods* **2000**, *5*, 343–355. [CrossRef] [PubMed]
50. Henson, R.K. Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha. *Meas. Eval. Couns. Dev.* **2001**, *34*, 177–189. [CrossRef]
51. Nimon, K.; Zientek, L.; Henson, R. The Assumption of a Reliable Instrument and Other Pitfalls to Avoid When Considering the Reliability of Data. *Front. Psychol.* **2012**, *3*, 102. [CrossRef] [PubMed]