A Comparison of Automated Scale Short Form Selection Strategies

Anthony W. Raborn
Pasco County Schools
7227 Land O' Lakes Blvd.
Land O' Lakes, FL 34638
1-727-774-4233
araborn@pasco.k12.fl.us

Walter L. Leite
University of Florida
1215 Norman Hall
Gainesville, FL, 32611
1-352-273-4302
Walter.leite@coe.ufl.edu

Katerina M. Marcoulides
University of Florida
1215 Norman Hall
Gainesville, FL, 32611
1-352-273-4332
k.marcoulides@coe.ufl.edu

ABSTRACT

Short forms of psychometric scales have been commonly used in educational and psychological research to reduce the burden of test administration. However, it is challenging to select items for a short form that preserve the validity and reliability of the scores of the original scale. This paper presents and evaluates multiple automated methods for scale short form creation based on metaheuristic optimization algorithms that incorporate validity criteria based on internal structure and relationships with other variables. The ant colony optimization (ACO) algorithm, tabu search (TS), simulated annealing (SA) and genetic algorithm (GA) are examined using confirmatory factor analysis (CFA) of scales with one factor, three factor, and bi-factor factorial structure. The results indicate that SA created short forms with best model fit for scales with one and three factor structures, but ACO was able to obtain highest reliability. For scales with bi-factor structure, SA provide short forms with best model fit, but TS obtained highest reliability. Overall, the SA algorithm is recommended because it produced consistently best model fit and reliability that was only slightly lower than the ACO or TS algorithms.

Keywords

Short form development, confirmatory factor analysis, metaheuristic algorithms, validity, ant colony optimization, tabu search, genetic algorithm, simulated annealing

1. INTRODUCTION

Applied researchers using psychometric scales often face a dilemma due to limited resources: should they use the full form of a well-established scale with strong validity evidence supporting it, but with a large number of items requiring a substantial amount of time and effort to complete, or should they use a short form of the scale that has not had the extensive evidence of validity? This issue has generated strong interest in the academic community the development of short forms of scales (e.g. [1]). Multiple methods have been proposed for scale short form development [2], with different fields utilizing a few preferred methods. For example, these methods include theoretical or practical justifications for the inclusion or exclusion of items (e.g., [3]), keeping one item from a set of items that are apparently similar or redundant (e.g., [4]), obtaining certain criteria for statistical values such as high factor loadings or item correlations (e.g., [5]), adding or retaining items that seem to improve measures of reliability and/or dimensionality (e.g., [6]).

The focus of item selection for short forms tends to be on the internal structure of the newly-created form, rather than using external relationships to help build the short form. For example, Petrillo, Capone, Caso, and Keyes [7] created a short form for a positive mental health assessment for use with Italian respondents by selecting items from twelve other scales with a focus on its internal structure. The resultant short form had adequate psychometric properties, but the average absolute correlation between the total score and sixteen other criterion measures was 0.37 (range: 0.20 to 0.62). Despite the adequate validity evidence for the internal structure, the external relationships would be characterized as modest since on average the short form’s and the other measures’ scores shared about 6% of their variances.

Obtaining a short form that has both adequate internal structure and strong validity with respect to relationships with other variables is difficult with traditional methods of short form development. Metaheuristic optimization algorithms [8] have the potential to solve these difficulties because the can simultaneously maximize multiple validity criteria for short forms. This paper aims to present the evaluation of multiple automated methods for short form creation based on metaheuristic optimization algorithms that incorporate criteria based on internal structure and relationships with other variables and determine which perform best under commonly used scale structures.

2. THEORETICAL FRAMEWORK

There have been some attempts to develop algorithms to derive short forms of scales that (a) maintain the internal structure of the scale in question (e.g. factor structure and/or content balance), (b) have favorable model characteristics such as meeting model fit statistic thresholds, and (c) produces scale scores that have favorable relationships with other variables, including other scales or external variables. For example, Olaru, Witthoft, and Wilhelm [9] compared multiple algorithms for the purpose of creating psychometrically valid short forms of a 99-item scale with various criterion (e.g., jointly optimizing two fit indices) and concluded that, under their study conditions, the Ant Colony Optimization (ACO) and Genetic Algorithm (GA) were able to produce statistically appropriate short forms that generalize well to new data. Marcoulides and Drezner [10] have shown that a Tabu search

Anthony Raborn, Walter Leite and Katerina Marcoulides "A Comparison of Automated Scale Short Form Selection Strategies" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 402 - 407

can be used to successfully reduce the number of items loading on factors. Leite, Huang and Marcoulides [2] developed and demonstrated an ACO algorithm that selects items for short forms while keeping adequate model fit and maximizing the relationship between the latent variable and external variables. More recently, Browne, Rockloff, and Rawat [11] produced an automated structural equation modeling (SEM) scale reduction algorithm and purport that it is an effective and efficient method for reducing items during scale development.

While these articles demonstrate the use of some automated scale short form development strategies and the importance of research in this area, an in-depth comparison of automated strategies under different scenarios does not seem to exist. In addition, some commonly-used metaheuristic algorithms for combinatorial problems have never been applied to the short-form development problem. For example, the inaugural example of the simulated annealing (SA) algorithm is with the Traveling Salesman Problem, which has various algorithms attempt to find the shortest path that travels between n cities exactly once [12], while no psychometric use of SA is apparent in the literature. To address these issues, this paper presents a simulation study utilizing three different scale structures commonly observed in educational research (one factor, three factor, and bifactor scales) and four meta-heuristic algorithms: The ant-colony optimization algorithm (ACO), genetic algorithm (GA), Tabu search (TS), and simulated annealing (SA). We chose these algorithms because they are the most well-established metaheuristic algorithms in the combinatorial optimization literature.

The ACO algorithm [13] mimics the behavior of ants searching for the shortest path to a food source. We evaluate the implementation of the ACO algorithm proposed by Leite, Huang and Marcoulides [2] for short form development, with minor modifications to the tuning parameters. Their implementation of the ACO algorithm attaches sampling weights to items, which are used to sample items for a set of candidate short forms of the scales. Each set of candidate short forms is evaluated and the best short form in the set is identified based on criteria that are specified by the researcher, such as SEM fit indices and the relationship between the scale's factors and an external variable. The criteria of choice are used to calculate the pheromone level, which is a summary of the quality of the short form chosen. The pheromone level is then used to update the sampling weights for the next round of sampling of candidate sets of short forms. This is repeated until a specified convergence criterion is met, such as number of iterations without improvement of the solution quality.

The GA mimics the biological process of evolution using the model parameters as genes. As implemented by Yarkoni [14], the algorithm generates an initial population of candidate models of size 200 and, by evaluating the fitness of each candidate model through a loss function, selects the best 20% of the models and repopulates. Between model generations, new models are created from mutation (randomly changing the items in a model) and recombination (two models exchanging items retained). After a certain number of iterations (100+), the model with the best fit according to the loss function is retained as the best solution. The loss function penalizes the fit of the models for every item included; this value needs to be tuned to achieve the correct reduction of items.

The TS algorithm implementation was modified from the presentation given in Marcoulides and Falk [15] to constrain the solution space to a specified number of items. Broadly, the TS looks at each of the local solutions to a model by changing one model parameter at a time; the particular change can be adjusted to suit the problem at hand. The main idea behind the TS procedure is to continually adjust the currently selected best model by examining other models in the neighborhood of the current best solution. If a neighboring model fits better than the current model, it is selected as the new best fitting model. If not, the examined neighboring model is marked "tabu"—placed on a list so that it is not reconsidered for some number of iterations. For this study, the TS was modified to (a) randomly generate a short form of a predetermined length from a longer form for the first iteration and (b) search for local short forms that maintain the predetermined length.

The SA algorithm is a statistical analog to metallurgic processes of annealing metals [16]. Generally, the algorithm begins with a specified starting model whose parameters are randomly changed by some process and a starting temperature. The new model is compared to the starting model and the difference in model fit is calculated. At any time, if the new model has better fit than the current model, it is selected for use in the next iteration; otherwise, the new model is selected with probability equal to a function of the difference in model fit and the current temperature. After each new model is either selected or ignored, the temperature updates and the current model is randomly changed. The algorithm checks each model against the best model seen and updates as needed. Some variants of the algorithm include a process that selects this best model after a certain number of iterations in which no better model has been found. This process repeats until the temperature reaches zero.

3. METHODS

3.1 Research Questions

1. How do the algorithms differ in terms of the time it takes for each to converge on a short form, model fit and reliability of the short form?

2. How do model misspecifications in the full form affect the fit and reliability of the short forms created by the algorithms?

3. Does the inclusion of an external variable affect the model fit and reliability of the short forms?

4. Does the performance of the algorithms depend of the factorial structure of the scale?

3.2 Manipulated conditions

To investigate the research questions, a Monte Carlo simulation study was conducted using the following population confirmatory factor analysis models: (a) the 20-item unidimensional model of the self-deceptive enhancement (SDE) scale [17], (b) the 24-item three-dimensional model of the teacher efficacy scale [18], and (c) a three-factor bifactor model [20] of the 30-item BASC-2 BESS [19] scale. These models represent three common models seen in scale development and are good representations of what educational researchers would work with. The covariance structure from the multidimensional scale were used to simulate samples for these conditions, and the factor loadings for the unidimensional model were used to simulate samples for this condition. In each case, the goal was to create a short form that is half the length of the long form.

Additional manipulated simulation conditions were the relationship with an external variable and full-scale model misspecification. For the relationship with an external variable, the two levels that were manipulated are (a) no relationship and (b) a moderate relationship (approximately equal to a path coefficient of 0.6 standard deviations). The full-scale model misspecification was manipulated in the simulation according to three levels: (a) no misspecification, (b) a minor misspecification in the factor loadings (i.e., population models modified to have six of the items cross-loading on a nuisance factor with a loading of 0.3), and (c) a major misspecification in the factor loadings (i.e., same as (b), but with factor loadings of 0.6 on a nuisance factor).

The data were simulated in R v3.5.0 [21] using the 'MASS' package [22]). The baseline condition (i.e., high reliability, no external variable relationship, no misspecification) used the values provided by the original models as the population values. These values were changed as necessary to create new population models that fit the target simulated conditions, resulting in fifteen covariance matrices. The sample size of each condition was set to 500. For each combination of manipulated conditions, we created 100 datasets.

3.3 Outcomes

The outcome variables of the simulation were the time to converge to a short form, the average level of model fit of the short form, and the composite reliability of the short form for each factor.

The comparative fit index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA) were used as the model fit indices, and the cutoff values of CFI > .95, TLI > .95, and RMSEA < .05 were used as indicators of adequate model fit [26].

For this study, the composite reliability of the one and three factor models was calculated as [23]:

$$CR_{\omega} = \frac{\left(\sum_{i=1}^{I} \lambda_i\right)^2}{\left(\sum_{i=1}^{I} \lambda_i\right)^2 + \sum_{i=1}^{I} \theta_i}$$

where the items are indexed with i, λ are the standardized factor loadings, and θ are the (standardized) residual variances of the items [24]. For the bifactor model, the composite reliability for the of general factor was calculated as

$$CR_{\omega_h} = \frac{\left(\sum_{i=1}^{k} \lambda_{g_i}\right)^2}{\left(\sum_{i=1}^{k} \lambda_{g_i}\right)^2 + \sum_{s=1}^{J} \left(\left(\sum_{i=1}^{k_s} \lambda_{si}\right)^2\right) + \sum_{i=1}^{k} \theta_i}$$

where s indexes the specific factors [25].

4. RESULTS

For each factor model, results for the "Minor Error with External Variable" condition did not have noticeable differences from either the "Minor Error with No External Variable" conditions or the "No Error" conditions, so this condition was dropped from the current study.

4.1 One Factor Model

The time to complete for each algorithm was similar across conditions, except for GA (see Table 1), which was faster. The time to converge was slightly longer for the ACO, SA, and TS under the major error with an external relationship condition.

The average model fit statistics for both factor models across 100 replications of the analysis for the three conditions is also shown in Table 1, where bolded values indicate good model fit. When there is no error in the original model, each algorithm produced good model fit, but as the error level increases the model fit decreased. In the major error conditions, only the SA algorithm had model fit greater than the traditional cutoff values.

Table 1. Model fit of short forms for one factor model

Error/ External	Method	Minutes to Complete	CFI	TLI	RMSEA
None/ No	ACO	2.482	**0.976**	**0.969**	**0.043**
	SA	2.516	**0.993**	**0.992**	**0.018**
	TS	3.933	**0.985**	**0.981**	**0.028**
	GA	0.699	**0.975**	**0.967**	**0.042**
Minor/ No	ACO	2.575	**0.961**	**0.950**	0.055
	SA	2.572	**0.987**	**0.983**	**0.027**
	TS	3.987	**0.977**	**0.971**	**0.036**
	GA	0.708	**0.964**	**0.953**	0.051
Major/ No	ACO	2.713	0.943	0.926	0.061
	SA	2.581	**0.983**	**0.978**	**0.029**
	TS	3.956	0.940	0.923	0.061
	GA	0.701	0.850	0.807	0.112
None/ Yes	ACO	3.059	0.943	0.929	0.058
	SA	2.871	**0.981**	**0.976**	**0.029**
	TS	2.819	0.923	0.903	0.066
	GA	0.702	0.859	0.823	0.106
Major/ Yes	ACO	3.301	0.942	0.928	0.058
	SA	3.107	**0.981**	**0.976**	**0.029**
	TS	5.162	0.934	0.917	0.060
	GA	0.752	0.855	0.819	0.108

The inclusion of the external variable reduced the model fit for each of the algorithms such that the conditions with no error and an external variable relationship had similar fit to the conditions with major error and either with or without external variable relationship (see Table 1).

Table 2 shows the reliability of the full form of the scale, as well as the reliability the short forms. As expected, the full form resulted in scores with higher reliability than all the short forms. The ACO had the greatest composite reliability for both the no error and minor error conditions, followed by the GA.

Table 2. Composite reliability estimates with one-factor model

Method	Reliability
Full Form	0.889
ACO	0.854
SA	0.813
TS	0.806
GA	0.835

4.2 Three Factor Model

For all five conditions, the SA and TS algorithm took about twice as long to converge on average as compared to the ACO and GA algorithms (see Table 3). For the three-factor model, the average model fit for the conditions with no external variable can be seen in Table 3. In both the no error and minor error conditions, each of the algorithms had good model fit. In the major error conditions, only the SA algorithm produced short forms with adequate model fit according to all three fit indices. The ACO and TS algorithms had good model fit according to CFI and TLI (ACO) and CFI (TS), while the GA algorithm had poor fit according to all three fit indices.

Table 3. Model fit for short forms with three-factor structure

Error/ External	Method	Minutes to complete	CFI	TLI	RMSEA
None/ No	ACO	2.967	**0.979**	**0.973**	**0.043**
	SA	6.228	**0.991**	**0.989**	**0.024**
	TS	6.063	**0.986**	**0.982**	**0.031**
	GA	2.705	**0.977**	**0.970**	**0.044**
Minor/ No	ACO	3.091	**0.978**	**0.971**	**0.047**
	SA	6.238	**0.989**	**0.985**	**0.030**
	TS	6.264	**0.984**	**0.979**	**0.037**
	GA	2.710	**0.974**	**0.967**	**0.048**
Major/ No	ACO	3.285	**0.964**	**0.953**	0.054
	SA	6.295	**0.990**	**0.987**	**0.027**
	TS	6.061	**0.956**	0.943	0.056
	GA	2.695	0.910	0.883	0.087
None/ Yes	ACO	3.713	**0.978**	**0.971**	**0.047**
	SA	7.040	**0.989**	**0.985**	**0.030**
	TS	7.577	**0.984**	**0.979**	**0.037**
	GA	2.685	**0.974**	**0.967**	**0.048**
Major/ Yes	ACO	3.387	**0.960**	**0.948**	0.060
	SA	6.173	**0.984**	**0.979**	**0.036**
	TS	6.586	**0.950**	0.935	0.063
	GA	2.282	0.917	0.892	0.087

Including an external variable had little effect on the model fit indices (see Table 3). With no error, the average model fit was approximately the same between the no external variable conditions and moderate external variable conditions, while the average model fit somewhat decreased in the major error condition for the external variable conditions as compared to the no external variable conditions. Only the SA produce short forms with good model fit across all the conditions.

The reliability of the full form and short forms with the three-factor CFA is shown in Table 4. All methods produced short forms with less reliable scores than the full form, but among the metaheuristic methods, the ACO produced short forms with the largest composite reliability for each of the factors in each condition.

Table 4. Composite reliability estimates with three-factor model

Method	Reliability Factor 1	Reliability Factor 2	Reliability Factor 3
Full form	0.870	0.910	0.900
ACO	0.788	0.846	0.833
SA	0.752	0.829	0.813
TS	0.754	0.828	0.809
GA	0.763	0.846	0.819

4.3 Bifactor Model

With the bifactor model, the GA had the fastest time to converge, and the ACO took about four times longer. The TS and SA algorithms had convergence times that were about 10 times of the GA algorithm.

Table 5 shows the average model fit indices of the bifactor model for the conditions with no external variable relationship. The SA and TS algorithms produced short forms with good model fit by each fit index in every condition, while the ACO resulted in good model fit by each fit index except for the RMSEA in the major error condition. The GA had good model fit by CFI in the no and minor error conditions only.

Including the external variable tended to reduce model fit. Both the SA and TS showed slight reductions in model fit across both error conditions, but still found short forms with good model fit according to all three fit indices. The ACO maintained approximately the same model fit in both no error and minor error conditions, but showed an increase in average fit in the major error conditions when comparing the no external variable to moderate external variable relationship conditions. However, only the CFI and TLI showed good model fit in these conditions (see Table 5).

The reliability of general factor with the full form and short forms with the bi-factor model are shown in Table 6. For the example scale used in this study, the full form produced scores with adequate reliability for the general factor, but for the specific factors the composite reliability is low. For the short forms, none of the algorithms produced consistently greater reliabilities for every factor in these conditions. The reliability of general factor with the short forms were smaller than the reliability of the full form for all algorithms. The GA performed best for the general factor reliability than the ACO, SA and TS.

Table 5. Model fit of short forms with bi-factor model and external variable

Error/ External	Method	Minutes to complete	CFI	TLI	RMSEA
None/ No	ACO	13.054	**0.986**	**0.976**	**0.041**
	SA	37.925	**0.995**	**0.992**	**0.020**
	TS	39.933	**0.992**	**0.986**	**0.028**
	GA	3.577	**0.951**	0.939	0.074
Minor/ No	ACO	12.296	**0.982**	**0.969**	**0.047**
	SA	40.658	**0.994**	**0.990**	**0.023**
	TS	48.200	**0.990**	**0.984**	**0.031**
	GA	3.562	**0.954**	0.943	0.069
Major/ No	ACO	12.875	**0.974**	**0.956**	0.055
	SA	39.556	**0.993**	**0.987**	**0.027**
	TS	41.211	**0.982**	**0.969**	**0.043**
	GA	3.606	0.945	0.921	0.079
None/ Yes	ACO	18.537	**0.986**	**0.976**	**0.041**
	SA	45.815	**0.995**	**0.992**	**0.020**
	TS	49.665	**0.992**	**0.986**	**0.028**
	GA	3.458	**0.951**	0.939	0.074
Major/ Yes	ACO	17.989	**0.975**	**0.959**	0.051
	SA	42.336	**0.991**	**0.986**	**0.028**
	TS	38.947	**0.979**	**0.965**	**0.046**
	GA	3.546	0.940	0.907	0.080

For the specific factors, the TS performed better than the other methods for two out of three factors. Surprisingly, the TS produced scores with higher reliability than the full form for factor 3, and the SA produced higher reliability than the full form for factor 2.

The results with the bi-factor model are limited in that the scale used produced scores with low reliability. Using a different scale that results in higher reliability of scores of the full form for all factors might have produced different results with respect to the comparison of algorithms.

Table 6. Reliability of short forms of general factor of bi-factor model

Method	General Factor	Factor 1	Factor 2	Factor 3
Full	0.715	0.304	0.114	0.319
ACO	0.661	0.067	0.115	0.062
SA	0.641	0.061	0.133	0.055
TS	0.659	0.149	0.124	0.433
GA	0.669	0.055	0.118	0.049

5. CONCLUSION

In general, the algorithms produced short forms with adequate model fit in all cases with no error, with two exceptions: each algorithm except the SA under the one factor model with an external variable, and the GA under the bifactor model. Therefore, when the original scale is correctly specified, the results showed that the algorithms are likely to produce short forms with model fit that maintain the desired factor structure of the scale.

The ACO, TS, and GA each had problems maintaining good model fit for the short forms with increasing error, though this was alleviated by increasing the factor structure's complexity. Including an external variable into the process generally had a small negative effect on average model fit, but the effect was never enough to cross the model fit thresholds. Overall, the SA provided short forms with the best average model fit in every single condition, while the ACO seemed to have better reliability on average for each of the factors with one factor and three-factor models. Given that the difference in reliability between the ACO and SA algorithms was about .05 or less on average, the practical difference in reliability between these methods may be outweighed by the difference model fit of the resulting short forms. Therefore, the current results lead to recommending the SA as the preferable metaheuristic algorithm for automated short form selection.

This study provides useful information to applied researchers about the benefits and drawbacks of utilizing these four algorithms for scale short form development in some common scenarios in educational research. This will allow for easier creation of psychometrically-sound short forms with stronger evidence of validity, especially as compared to creating short forms manually. Future research could apply these algorithms to the short form creation problem alongside other methods on real data to compare the efficacy of each approach.

6. REFERENCES

[1] Fischer, D.G. and C. Fick, *Measuring Social Desirability: Short Forms of the Marlowe-Crowne Social Desirability Scale.* Educational and Psychological Measurement, 1993. **53**(2): p. 417-424.

[2] Leite, W.L., I.C. Huang, and G.A. Marcoulides, *Item Selection for the Development of Short Forms of Scales Using an Ant Colony Optimization Algorithm.* Multivariate Behavioral Research, 2008. **43**(3): p. 411-431.

[3] Noble, W., et al., *A short form of the Speech, Spatial and Qualities of Hearing scale suitable for clinical use: The SSQ12.* International journal of audiology, 2013. **52**(6): p. 409-412.

[4] Dennis, C.L., *The breastfeeding self-efficacy scale: Psychometric assessment of the short form.* Journal of Obstetric, Gynecologic & Neonatal Nursing, 2003. **32**(6): p. 734-744.

[5] Wester, S.R., et al., *Development and evaluation of the Gender Role Conflict Scale Short Form (GRCS-SF).* Psychology of Men & Masculinity, 2012. **13**(2): p. 199.

[6] Lim, S.Y. and E. Chapman, *Development of a short form of the attitudes toward mathematics inventory.* Educational Studies in Mathematics, 2013. **82**(1): p. 145-164.

[7] Petrillo, G., et al., *The Mental Health Continuum–Short Form (MHC–SF) as a measure of well-being in the Italian context.* Social Indicators Research, 2015. **121**(1): p. 291-312.

[8] Dréo, J., et al., *Metaheuristics for hard optimization*. 2006, New York: Springer.

[9] Olaru, G., M. Witthöft, and O. Wilhelm, *Methods matter: Testing competing models for designing short-scale Big-Five assessments.* Journal of Research in Personality, 2015. **59**: p. 56-68.

[10] Marcoulides, G.A. and Z. Drezner, *Tabu search variable selection with resource constraints.* Communications in Statistics: Simulation & Computation, 2004. **33**(2): p. 355-362.

[11] Browne, M., M. Rockloff, and V. Rawat, *An SEM Algorithm for Scale Reduction Incorporating Evaluation of Multiple Psychometric Criteria.* Sociological Methods & Research, 2018. **47**(4): p. 812–836.

[12] Kirkpatrick, S., C. Gelatt, and M. Vecchi, *Optimization by simulated annealing.* Science, 1983. **220**(4598): p. 671-680.

[13] Colorni, A., M. Dorigo, and V. Maniezzo, *Distributed optimization by ant colonies*, in *Proceedings of ECAL 91 - First European Confernce on Artificial Life*, F. Varela and P. Bourgine, Editors. 1992, Elsevier Publishing: Paris, France.

[14] Yarkoni, T., *The abbreviation of personality, or how to measure 200 personality scales with 200 items.* Journal of Research in Personality, 2010. **44**: p. 180-198.

[15] Marcoulides, K.M. and C.F. Falk, *Model Specification Searches in Structural Equation Modeling with R.* Structural Equation Modeling: A Multidisciplinary Journal, 2018. **25**(3): p. 484-491.

[16] Drezner, Z., G.A. Marcoulides, and S. Salhi, *Tabu search model selection in multiple regression analysis.* Communications in Statistics - Simulation and Computation, 1999. **28**(2): p. 349-367.

[17] Leite, W.L. and S.N. Beretvas, *Validation of Scores on the Marlowe-Crowne Social Desirability Scale and the Balanced Inventory of Desirable Responding.* Educational and Psychological Measurement, 2005. **65**(1): p. 140-154.

[18] Tschannen-Moran, M. and A.W. Hoy, *Teacher efficacy: Capturing an elusive construct.* Teaching and teacher education, 2001. **17**(7): p. 783-805.

[19] Reynolds, C.R., R.W. Kamphaus, and K.J. Vannest, *Behavior assessment system for children (BASC)*, in *Encyclopedia of clinical neuropsychology*. 2011, Springer: New York. p. 366-371.

[20] Splett, J.W., et al., *Factor analytic replication and model comparison of the BASC-2 Behavioral and Emotional Screening System.* Psychological assessment, 2017. **29**(12): p. 1543.

[21] R Development Core Team, *R: A language and environment for statistical computing*. 2018, R Foundation for Statistical Computing: Vienna, Austria.

[22] Venables, W.N. and B.D. Ripley, *Modern applied statistics with S*. 2002, New York:: Springer Science & Business Media,.

[23] Raykov, T., *Estimation of Composite Reliability for Congeneric Measures.* Applied Psychological Measurement, 1997. **21**(2): p. 173-184.

[24] McDonald, R.P., *Test theory: A unified treatment*. 1999, Mahwah, NJ: Lawrence Erlbaum.

[25] Gignac, G.E. and M.W. Watkins, *Bifactor modeling and the estimation of model-based reliability in the WAIS-IV.* Multivariate Behavioral Research, 2013. **48**(5): p. 639-662.

[26] Hu, L. and P.M. Bentler, *Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives.* Structural Equation Modeling, 1999. **6**: p. 1-55.