Editorial

Are the Machines Taking Over?

Benefits and Challenges of Using Algorithms in (Short) Scale Construction

Jan Dörendahl and Samuel Greiff

Cognitive Science and Assessment (COSA), University of Luxembourg, Luxembourg

Technological advancements have changed the way we construct psychological assessment tools. Scale construction algorithms are automated item selection procedures that were developed to refine large numbers of items into valid scales to assess psychological constructs. So far, this technology has mostly been used to shorten established scales – for example, in order to meet time and monetary restrictions in large-scale assessments (Rammstedt & Beierlein, 2014). However, the principle of algorithms can be transferred to other scenarios, such as refining newly developed item pools into scales when constructing a new assessment tool from scratch. As with all new technologies, the question arises whether algorithm-based item selection is just a fancy gimmick or in fact a powerful alternative to traditional approaches for improving psychological assessment tools. In this editorial, we would like to shed some light on the pros and cons of using algorithms in (short) scale construction.

Challenges in Scale Development

State-of-the-art scale development poses a significant challenge for scale developers, who must find a combination of suitable items from a (sometimes very) large item pool. This combination of items (1) must be capable of assessing the trait the researcher or practitioner is interested in, while also (2) ensuring reliable assessment and resulting in (3) a scale with high validity. For instance, construct validity requires that the scale has a solid internal structure. This can be empirically demonstrated through a confirmatory factor analysis (CFA) with a good model fit according to fit indices such as Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA). The following hypothetical example using real data shows that this challenge can be a tough one indeed.

In motive research, explicit motives are typically assessed using self-report questionnaires (McClelland, Koestner, & Weinberger, 1989). In our own research, as part of constructing a scale to assess one such explicit motive, the autonomy motive (i.e., need to be free from others' influence and interference; Patrick, Skinner, & Connell, 1993), we developed a 29-item pool with the goal of developing a scale that would ultimately consist of nine items.[1] We wanted the scale to demonstrate good fit in a CFA, with an RMSEA as low as possible and a CFI as high as possible. Given that the order of items in a CFA does not matter for calculating model fit and each item may only appear once among the nine selected items, this particular item selection scenario (i.e., 9 out of 29 items) encompassed a total of 10,015,005 possible combinations. Closely examining all possible combinations is impossible by hand and certainly not efficient.

Algorithms such as ant colony optimization (ACO; Marcoulides & Drezner, 2003) can serve as a short cut to achieving this aim, as they are able to process large numbers of items and compare the scales constructed from them to multiple prespecified criteria such as CFA model fit, reliability, and correlations with external criteria. This facilitates the automatized optimization of several criteria in scale development. In comparison, selecting items based just on common sense and content sometimes fails to consider empirical support for the scale's construct validity, for instance in terms of CFA results, and to optimize the scale's reliability. Another rather simple procedure that is commonly used, particularly in short scale construction is

[1]The authors would like to express their gratitude to the many different people that were involved in generating items, contributing ideas, and participating in helpful discussions. Not only the authors of this Editorial, but an entire group of people was involved in generating the initial item pool.

https://doi.org/10.1027/1015-5759/a000597

maximizing the main factor loadings, although this approach has limitations as well. As the items are selected based on their loadings in an EFA or CFA, this procedure results in a highly reliable scale. However, the resulting scale often subsequently fails to demonstrate acceptable model fit when tested in a CFA (Olaru, Witthöft, & Wilhelm, 2015).

Advantages of Algorithms in (Short) Scale Construction

The ACO algorithm mentioned above is relatively easy to implement and seems to outperform other algorithms and manual item selection approaches, such as maximizing main loadings (Leite, Huang, & Marcoulides, 2008; Olaru et al., 2015). For this editorial, it serves as an example as it provides a transparent, empirically driven optimization procedure based on criteria chosen by the scale developer. The algorithm is based on the behavior of ants searching for food (Deneubourg, Pasteels, & Verhaeghe, 1983), who leave a trail of pheromones that attracts other ants. At first, the ants will traverse random routes between the formicary and the food source. However, the journey takes less time on shorter routes, meaning that pheromones accumulate faster on shorter compared to longer routes. Consequently, a growing number of ants are attracted to the shortest route, optimizing the path between the formicary and food source. ACO functions in a similar way in item selection, as it optimizes the selection of a set of items into a scale based on prespecified criteria. The algorithm works with probabilities to save time and computational effort compared to computing all possible solutions (i.e., brute force search or exhaustive search; e.g., Leite et al., 2008).

Returning to our item selection example, we wanted the final scale for the autonomy motive to fulfill strict measurement criteria. Accordingly, we selected CFA model fit in terms of CFI and RMSEA as the optimization criteria. ACO first selects several random sets of items and compares the scales built from these items to the prespecified selection criteria (i.e., a high CFI and a low RMSEA). Based on the results, ACO then increases the probability of items that perform better with respect to these criteria being selected in future solutions. Over the course of up to several thousand iterations, the algorithm iteratively adjusts the items' probability of being selected until a particular combination of autonomy items cannot be further improved upon. This combination of items is then presented as the final autonomy scale. In contrast, brute force search would have involved estimating all 10,015,005 possible combinations and checking their adherence to the prespecified criteria. In the provided example, however, ACO only had to estimate 10,619 models to identify an autonomy scale with perfect fit indices (i.e., CFI = 1.00; RMSEA = 0.00). Compared to brute force search, it took ACO only 0.11% of the computational effort to find an autonomy scale fulfilling the specified criteria. While we chose model fit in terms of CFI and RMSEA as the optimization criteria in this example, the algorithm makes it possible to optimize any aspect of the scale, such as reliability or correlation with an external criterion. From a theoretical perspective, there is no limit to how many optimization criteria can be combined. For example, Schroeders, Wilhelm, and Olaru (2016a) combined model fit, reliability, sensitivity, and correlations with covariates in an item selection process using ACO.

Disadvantages of Algorithms in (Short) Scale Construction

Algorithms such as ACO can eliminate a lot of the cumbersome work in scale development and are powerful tools for constructing (short) scales of high psychometric quality from large item pools with relatively low effort. However, algorithms are obviously limited to the quality of the item pool from which they select (i.e., garbage in, garbage out).

If the item pool contains only low-quality items, the result will not be a psychometrically sound scale.

Moreover, ACO requires the specification of quantitative optimization criteria, such as model fit coefficients, correlation coefficients with external criteria, reliability coefficients, or measurement invariance criteria. Put differently, the algorithm is "blind" to content coverage and does not check whether all theoretically relevant aspects of the construct are adequately represented in the final scale. This is, obviously, a major shortcoming of automatized algorithms in general. Consequently, the scale developer needs to once again take an active role at this point and examine the scale from a theoretical perspective. In addition, because the algorithm optimizes an item pool into a scale on the basis of a specific sample, overfitting may occur (Olderbak et al., 2015). Consequently, the psychometric properties of a scale constructed with ACO might be somewhat different when tested in a sample other than the construction sample.

In sum, there are several aspects to consider when working with ACO in order to achieve the desired optimization. Specifically, (1) make sure the item pool fully covers the construct from a theoretical point of view (regardless of whether it has been developed from scratch or whether it is a long scale that needs to be shortened), (2) check whether the final scale identified by the algorithm contains all key aspects of the construct and make manual modifications if necessary,

and (3) cross-validate the results in a different sample. If you follow these guidelines, algorithms such as ACO can help you choose a good set of items for your scale.

Do Use Algorithms in Scale Development, But. . .

As this example regarding ACO demonstrates, algorithms are useful in test construction and can eliminate some of the rather cumbersome parts. However, they do not reduce the knowledge and skills needed on behalf of the scale developer – in fact, the opposite is the case. Researchers need to have profound knowledge of psychometric properties and the specification of measurement models in CFA and must choose the optimization criteria wisely. In addition to such statistical knowledge, a solid understanding of the construct to be assessed is indispensable in order to evaluate the scales produced by the algorithms. If the scale developer understands what the algorithm is both capable of and, equally important, not capable of, it can be one powerful tool (out of several) that we recommend using in (short) scale construction.

In *EJPA*, we have not often seen papers that employ ACO or any other algorithm. Of course, scale construction will never solely rely on automatic algorithms (and for good reasons as outlined above), but they can be an interesting addition to existing procedures and they have the potential to add further value to papers published in *EJPA*. If – through the example above – the use of algorithms in (short) scale construction has peaked your interest and you would like to learn more, we recommend Janssen, Schultze, and Grötsch (2017), Olaru et al. (2015), and Schroeders et al. (2016a) for introductions to and comparisons of different algorithms and more traditional methods of short scale development as well as examples of complex optimization criteria. For an advanced application of ACO in multi-group structural equation modelling, the interested reader may consult Schroeders, Wilhelm, and Olaru (2016b). And if these readings further convinces you, go ahead and check whether your research might benefit from using any of these algorithms.

References

Deneubourg, J. L., Pasteels, J. M., & Verhaeghe, J. C. (1983). Probabilistic behaviour in ants: A strategy of errors? *Journal of Theoretical Biology, 105*(2), 259–271. https://doi.org/10.1016/S0022-5193(83)80007-1

Janssen, A. B., Schultze, M., & Grötsch, A. (2017). Following the ants: Development of short scales for proactive personality and supervisor support by ant colony optimization. *European Journal of Psychological Assessment, 33*, 409–421. https://doi.org/10.1027/1015-5759/a000299

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research, 43*(3), 411–431. https://doi.org/10.1080/00273170802285743

Marcoulides, G. A., & Drezner, Z. (2003). Model specification searches using ant colony optimization algorithms. *Structural Equation Modeling: A Multidisciplinary Journal, 10*(1), 154–164. https://doi.org/10.1207/S15328007SEM1001_8

McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review, 96*(4), 690–702. https://doi.org/10.1037/0033-295X.96.4.690

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality, 59*, 56–68. https://doi.org/10.1016/j.jrp.2015.09.001

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: Toward a brief form for research and applied settings. *Frontiers in Psychology, 6*, 1–14. https://doi.org/10.3389/fpsyg.2015.01503

Patrick, B. C., Skinner, E. A., & Connell, J. P. (1993). What motivates children's behavior and emotion? Joint effects of perceived control and autonomy in the academic domain. *Journal of Personality and Social Psychology, 65*(4), 781–791. https://doi.org/10.1037/0022-3514.65.4.781

Rammstedt, B., & Beierlein, C. (2014). Can't we make it any shorter? The limits of personality assessment and ways to overcome them. *Journal of Individual Differences, 35*(4), 212–220. https://doi.org/10.1027/1614-0001/a000141

Schroeders, U., Wilhelm, O., & Olaru, G. (2016a). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLoS One, 11*(11), e0167110. https://doi.org/10.1371/journal.pone.0167110

Schroeders, U., Wilhelm, O., & Olaru, G. (2016b). The influence of item sampling on sex differences in knowledge tests. *Intelligence, 58*, 22–32. https://doi.org/10.1016/j.intell.2016.06.003

Published online April X, 2020

Jan Dörendahl
Cognitive Science and Assessment (COSA)
University of Luxembourg
2, avenue de l'Université
4365 Esch sur Alzette
Luxembourg
jan.dorendahl@uni.lu

Samuel Greiff
Cognitive Science and Assessment (COSA)
University of Luxembourg
2, avenue de l'Université
4365 Esch sur Alzette
Luxembourg
samuel.greiff@uni.lu