Balancing Between Categorical and Dimensional Assessment in Short-Scale Construction Using Ant Colony Optimization

Priscilla Achaa-Amankwaa[1], Tim Trautwein[1], Wolfgang Lenhard[2], Ulrich Schroeders[1]

[1] Department of Psychology, University of Kassel, Germany

[2] Department of Psychology IV, University of Würzburg, Germany

Author Note

Priscilla Achaa-Amankwaa https://orcid.org/0000-0003-4087-5293

Tim Trautwein https://orcid.org/0009-0008-3443-1876

Wolfgang Lenhard https://orcid.org/0000-0002-8184-6889

Ulrich Schroeders https://orcid.org/0000-0002-5225-1122

We have no conflict of interest to disclose. We confirm that the work conforms to Standard 8 of the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct. A synthesized data set, a results data table, and R-syntax to reproduce the findings presented here are available at https://osf.io/pe5sy/?view_only=33f584b43bbe4dfb873c690591165576 (Achaa-Amankwaa et al., 2024). The Ant Colony Optimization algorithm is provided with annotated syntax on GitHub for experimentation and replication purposes: https://github.com/UnknownBonobo/ACO-Cat-vs.-ACO-Dim.

Correspondence concerning this article should be addressed to Priscilla Achaa-Amankwaa, Department of Psychology, University of Kassel, Holländische Straße 36-38, 34127 Kassel, Germany. E-Mail: Priscilla.achaa@uni-kassel.de

Abstract

Language proficiency assessment poses particular challenges for test developers in selecting items that allow for a clear assignment of individuals to language proficiency levels (categorical assessment), while at the same time providing a reliable and comprehensive dimensional assessment of language proficiency. We show how Ant Colony Optimization (ACO) can be used to achieve a balance between these measurement goals, using a German entry-level language assessment as a working example. We tailored competing ACO algorithms to develop short scales of different lengths that met several pre-specified criteria, including model fit, composite reliability, and criterion validity. In optimizing the short scales, we favored either accurate dimensional assessment (model fit and composite reliability), between-category classification accuracy (a high polychoric correlation between model-predicted and independently assessed proficiency levels), or a balance of both. We argue that scale optimization strategies such as ACO are essential for balancing conflicting measurement goals such as optimizing between categorical and dimensional assessment.

Keywords: Ant Colony Optimization, metaheuristics, categorical assessment, dimensional assessment, language assessment

Balancing Between Categorical and Dimensional Assessment in Short-Scale Construction Using Ant Colony Optimization

Developing reliable and efficient measurement scales is fundamental in psychological assessment. Measurement scales with many items have served as a standard for collecting reliable and valid data for a long time. However, they are not optimally suited for many current applications: In multivariate, intense longitudinal, and large-scale studies, lengthy scales bear the risk of participant fatigue or dropout, which is why there is an increasing demand for shorter, yet equally psychometrically sound measurement scales. However, a significant reduction in test length usually comes at the expense of reliability and validity (Kruyen et al., 2013). Thus, test developers are often faced with the challenging task of selecting those specific items from a larger item pool that still allow adequately reliable and valid measurement. Innovative approaches, including metaheuristic optimization techniques such as Ant Colony Optimization (ACO), help compile psychometrically sound short scales (Leite et al., 2008; Schroeders et al., 2016a; Olaru, Schroeders, Hartung, & Wilhelm, 2019). In the present study, we extend previous applications of ACO to construct reliable and valid short scales that can be used in ordered categorical assessment. We use an existing German entry-level language assessment as a working example of how to set up the optimization procedure to arrive at a highly homogeneous and reliable short scale that simultaneously includes items that can accurately differentiate between proficiency levels.

Categorical Assessment of Dimensionally Distributed Constructs

Categorical assessment has a long tradition in clinical psychology (Kamphuis & Noordhof, 2009) and educational psychology (Hickendorff et al., 2018; Van der Maas & Molenaar, 1992). In language assessment, categorical assessment is often used to group individuals into language proficiency levels in alignment with a national or international competency framework (e.g., Papageorgiou, 2016). The statistical and methodological problems associated with categorizing, or even dichotomizing, a continuous scale are identical across applications and have been repeatedly highlighted (e.g., Irwin & McClelland, 2003; MacCallum et al., 2002). For example, methodological research has shown that artificial categorization of continuous variables—such as building groups to reflect clinical symptom severity, academic competencies, or language proficiency levels based on arbitrary cutoffs—usually results in information loss about the individual variation on the variable, potentially obscuring important differences in the relationships between variables. This in turn contributes to biased effect sizes (under- or overestimation), spurious effects, and the potential to overlook complex nonlinear relationships (MacCallum et al., 2002). Moreover, the categorization of variables that are continuous in nature leads to a loss in statistical power and measurement precision (Bennette & Vickers, 2012; McClelland et al., 2015).

Notwithstanding these concerns, categorical assessment continues to enjoy great popularity in practice. One reason for this is that reducing the amount of information facilitates the communication of assessment procedures and results to non-experts (DeCoster et al., 2011). In addition, test results are often used to make categorical decisions, such as whether patients should be treated with psychotropic medication or what type of educational intervention is most appropriate in an educational setting. Since categorical assessment is often desired in practice, the measurement intention is to optimally differentiate between groups. Test developers then often adopt a criterion-related or external test construction strategy, by which the test items are chosen to maximize the discriminative accuracy regarding group or category membership. In contrast to internal test construction strategies, in which item selection often results in maximizing homogeneity (Steger et al., 2023), external strategies select items that tend to maximize heterogeneity between groups.

Short-Scale Development Using Meta-Heuristics

Traditional short-scale construction usually takes a single optimization criterion into account, for example, selecting items based on some item-level statistics such as the item-total correlation or the alpha-if-item-deleted statistic (for an overview see Kruyen et al., 2012, 2013). The focus of item selection is often on reliability, as validity issues are much more difficult to define and address (Markus & Borsboom, 2013). This often leads to reduced heterogeneity in the item pool, that is, lower construct coverage. Furthermore, in the process of scale shortening, test developers tend to eliminate items stepwise. Once an item is removed, subsequent decisions are made based on the remaining item set, potentially overlooking better configurations due to a limited exploration of the solution space (Olaru et al., 2015; Schroeders et al., 2016b). Prominent metaheuristics that rely on nature-inspired processes to solve complex combinatorial problems include Ant Colony Optimization (Leite et al., 2008), the Genetic Algorithm (Eisenbarth et al., 2015; Yarkoni, 2010), Tabu Search (Glover, 1986; Marcoulides & Falk, 2018), and Simulated Annealing (Černý, 1985; Drezner & Marcoulides, 1999). These metaheuristics have repeatedly been shown to outperform the above-mentioned traditional item selection methods in short-scale construction, particularly because they offer the possibility to consider multiple criteria simultaneously in an automated and computationally efficient search for optimal solutions (for an overview and comparison, see Marcoulides & Ing, 2012; Raborn et al., 2020). In addition, new algorithms inspired by natural phenomena, such as the foraging behavior of bees (Schroeders et al., 2024) are continually being introduced into test development and structural equation modeling.

In psychological test development, ACO has been one of the earliest and most widely used optimization algorithms for developing and optimizing short scales (e.g., Janssen et al., 2017; Leite et al., 2008; Schroeders et al., 2016a, 2016b; Olaru, Schroeders, Hartung, & Wilhelm, 2019). The various software packages for ACO (e.g., *ShortForm* by Raborn &

Leite, 2018, or *stuart* by Schulze, 2023), make it easy and convenient to implement. The algorithm is inspired by the foraging behavior of ant colonies. In an initial exploration of the environment around the nest, ants randomly leave pheromone trails. On shorter paths to the food source, more pheromones accumulate over time, attracting more and more ants, thus, reinforcing the trails, whereas the pheromone trails of longer or less promising paths evaporate with time. Similar to ants leaving pheromones along the shortest route between the nest and a food source, ACO uses virtual pheromones to enhance the attractiveness of item sets that better meet predefined psychometric properties during scale construction (e.g., model fit and reliability). The pheromones in the biological world equate to the drawing probabilities in the item sampling procedure. Across several iterations, the item sets are evaluated according to the predefined optimization criteria, and the drawing probabilities are updated based on the solution quality. The endpoint of this iterative and non-greedy search process is typically a near-to-optimal solution that suffices all criteria but is not necessarily the best solution due to the probabilistic nature of ACO. Choosing a high number of ants, a high number of maximum iterations, and a low evaporation rate, as well as running the algorithm multiple times with different seeds, is therefore recommended to ensure a good balance between exploration and exploitation of model solutions, but at the cost of increased computation time (Olaru, Schroeders, Hartung, & Wilhelm, 2019).

The algorithm’s versatility and ability to handle complex optimization problems make it well-suited for the various requirements of psychological assessment. Short-scale construction using ACO has been, for example, implemented to optimize knowledge assessment by targeting item sets that are reliable and still cover the breadth of the multidimensional construct (Steger et al., 2023) or to construct parallel test forms in an item response framework (Zimny et al., 2024). In the realm of personality assessment, ACO has been used to compile personality scales that exhibit good model fit, high reliability, and content validity (e.g., Janssen et al., 2017). Moreover, these short scales can be designed to be measurement invariant across countries (Jankowsky et al., 2020; Olaru & Danner, 2021) or age (Olaru et al., 2018; Olaru & Jankowsky, 2022). ACO has also been used to amplify differences between groups to learn more about the influence of item sampling on the construct, specifically age differences in personality (Olaru, Schroeders, Wilhelm, & Ostendorf, 2019), and gender differences in declarative knowledge (Schroeders et al., 2016b). Previous implementations of ACO have focused on the multi-criteria optimization of dimensional models. As ACO has not yet been applied to optimize the accurate assignment to ordinal categories, this study demonstrates the first use of ACO for this objective.

The Present Study

In the present study, we aim to extend the implementation of ACO to the optimization of language assessment. To this end, we demonstrate the method's applicability as a proof-of-concept by constructing several short scales of the E-DaF placement test for German as a foreign language (Einstufungstest für Deutsch als Fremdsprache; Lenhard et al., 2024). The test has 102 items and provides an economic classification of general language proficiency in German for levels A1 to B2 of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). The aim of the short-scale construction is, on the one hand, to maximize the accuracy of classification into proficiency levels (categorical assessment) and, on the other hand, to achieve a good model fit of a unidimensional model where the latent variable is also measured reliably (dimensional assessment). As we explain in more detail in the next paragraph, these two aims are hard to reconcile in a single modeling framework because the items that contribute to a fine-grained measurement of language proficiency in a reflective measurement model may not always aid in distinguishing between levels where a formative measurement approach might be a better choice, and vice versa.

Language proficiency, as measured by the E-DaF, follows a unidimensional normal distribution. When individuals are classified into CEFR proficiency levels, this dimensional distribution is discretized into ordered categories, wherein the boundaries between categories remain blurred. Therefore, to optimize accurate classification in language assessment, the overlap between the proficiency levels should be minimized, which can be achieved by selecting items whose difficulty varies between observed proficiency levels. This approach aligns more closely with formative assessment within a regression framework, where test items are selected based on their contribution to differentiating between groups or predicting an external criterion. Conversely, in reflective measurement within a structural equation modeling framework, the items are manifestations of an underlying latent trait, and thus, internal consistency and model fit are prioritized. An item that strongly differentiates between two proficiency levels might not be ideal for capturing more nuanced ability differences. To achieve the best of both worlds in scale optimization, it is therefore essential to consider both approaches to meet these two conflicting objectives.

Addressing this optimization problem, the goal of the present study is to construct a short scale of the E-DaF that strikes a balance between the optimal discrimination of different language proficiency levels and a reliable measurement model of language proficiency. We tailored three ACO algorithms to optimize the following criteria: a) incremental and absolute model fit of a unidimensional factor model as well as the factor saturation (i.e., composite reliability), b) the polychoric correlation between the predicted proficiency levels and the observed levels assessed independently of the actual test, and c) a combination of both to possibly arrive at a balance between the test construction objectives. The factor saturation reflects the measurement accuracy of language proficiency. The polychoric correlation serves as an indicator of criterion validity because it describes the relationship between the proficiency level based on the test score and a categorical, external language proficiency assessment. A higher correlation reflects a greater test accuracy in assigning persons to proficiency levels. We constructed three versions of the optimization function with differing focuses on model fit and factor saturation and the polychoric correlation . To find the optimal length of the short scale, we ran each optimization algorithm across 4 and 20 selected items.

Method

Participants and Instrument

The analysis sample included 581 participants (56% female) from the standardization sample of the E-DaF (Lenhard et al., 2024), which ranks language proficiency from levels A1 (beginner) to B2 (upper intermediate) according to the CEFR. Participant ages ranged from 16 to 47 years, with a mean age of 24.09 years (*SD* = 4.95). Placement into language proficiency levels was based on an external language assessment, either through language certificates obtained in the participants' country of origin or introductory language courses in Germany. A total of 96 participants (17%) who were enrolled in but had not yet completed an A1 course were assigned to the pre-A1 level. The remaining assignment was as follows: 128 participants (22%) were ranked at level A1, 87 (15%) at level A2, 153 (26%) at level B1, and 117 (20%) at level B2. For all analyses, participants at the pre-A1 and A1 levels were combined, as the CEFR framework begins at level A1. The participant's native languages were placed in the following language groups: Romance languages (180, Spanish, Portuguese, Italian, etc.), Germanic languages (62, mostly English), Slavic languages (46, Russian, Polish, etc.), Indo-Iranian languages (40; Persian, Dari, Pashto, etc.), Arabic (37), Korean (37), Turkic languages (33), Chinese (25), Japanese (20), and other languages (47). 53 people did not provide any information on their native language.

The E-DaF is designed for use in the language proficiency assessment for vocational training and qualification measures, the selection of suitable language courses, and personnel selection. The scope of application thus ranges from formative assessment in low-stakes settings to high-stakes assessments for language certification. It is tailored to adolescents and adults with non-German native language who are learning German as a foreign language and have recently moved to a German-speaking country or plan to move there. The long form with 102 single-choice items (one correct answer out of four), assesses receptive language skills in the three central language aspects lexis, grammar, and orthography. A comparative study with the widely used German online placement test onSet (Eckes, 2014) showed an 86% agreement in proficiency assessments among 49 students with non-German language backgrounds.

Analyses

For the dimensional assessment, we estimated a unidimensional confirmatory factor analysis using the Weighted Least Squares Mean and Variance adjusted (WLSMV) estimator. We excluded all extremely easy items (P > .97) to avoid convergence issues caused by low item variance. This preparational step reduced the item set to 99 items.

The ACO script is an extended version of a script provided by Leite (2008). For a more detailed description of ACO, we recommend Dorigo and Stützle (2010) and Olaru, Schroeders, Hartung, & Wilhelm (2019). We provide an annotated syntax for a more detailed description of the ACO algorithm on GitHub: https://github.com/UnknownBonobo/ACO-Cat-vs.-ACO-Dim). At the core of the ACO algorithm is an optimization function that evaluates multiple criteria such as reliability and model fit. To standardize the criteria on a common scale ranging from 0 to 1, all values were logit-transformed (see Appendix, formulae (1) – (4)). Item sets with values close to or better than the specified cut-off criteria receive higher pheromone values (φ). Specifically, we evaluated 1) the CFI, 2) the RMSEA, 3) the factor saturation, that is, composite reliability McDonald’s ω, and 4) the classification accuracy, which we operationalized as the polychoric correlation between the model-predicted and the observed (independently assessed) proficiency levels (r_{polychor}). The prediction was performed using ordinal logistic regression, assuming a cumulative logit model (see Agresti, 2019) with the external ordered language proficiency levels as the dependent variable. Importantly, this model assumes that the regression coefficients, equivalent to the log odds ratios for the predictors in a logit model, are the same across all thresholds of the ordinal outcome. This is referred to as proportional odds assumption (Agresti, 2019). We tested this assumption using Brant tests for each regression model. For any item set that failed this test with $\alpha \leq .05$, the pheromone values were set to 0, and these sets were excluded from further consideration in the ACO process to ensure that only item sets that met the proportional odds assumption were optimized.

To ensure model fit, we set the cutoff criteria for good model fit to CFI ≥ .98 and RMSEA ≤ .02 (see Yu, 2002 for recommended cutoff values for categorical indicators). For the factor saturation, we chose a cutoff value of McDonald's $\omega \geq .80$, which is deemed acceptable for exploratory research or low-stakes assessments (Nunnally & Bernstein, 1994). Lastly, we set the cutoff for $r_{\text{polychor}} \geq .90$.

We ran three separate ACO procedures taking the above criteria into account (see Appendix, formulae (5) – (7)). Optimizing dimensional assessment (ACO_{Dim}), the first procedure optimized model fit of a unidimensional confirmatory factor model and McDonald's ω, that is, reliable measurement of proficiency levels. Optimizing categorical assessment (ACO_{Cat}), the second procedure focused on optimizing r_{polychor} solely as an indicator of the predictive accuracy of language level classifications. With the third procedure, ACO_{Bal}, we strived to achieve good model fit and a balance between McDonald's ω and r_{polychor} by weighting both criteria equally high in the optimization function. We ran each of the three ACO optimization procedures across 5 seeds to enhance the robustness of results and to reduce the likelihood of capitalizing on chance, and selected the best-performing solution for each seed. The analyses were performed across 11 different scale lengths (4 to 10, 12, 15, 18, 20 items) with the following hyperparameters: 120 ants, a maximum of 90 iterations, and an evaporation rate of .99 (with higher values indicating slower evaporation, encouraging exploitation of known paths).

All analyses were conducted in R (version 4.2.1; R Core Team, 2022), the full analysis code—including the wrapper function—of the ACO algorithm is published on GitHub (https://github.com/UnknownBonobo/ACO-Cat-vs.-ACO-Dim). The results table of our ACO analysis and R-syntax to reproduce the figures and tables presented here are available via OSF at https://osf.io/pe5sy/?view_only=86a967a28bbb4b8c8f2f946af93c03dd. As the E-DaF data of the standardization sample are subject to publisher copyright, we provide a synthesized dataset that was created with *synthpop* (version 1.8-0; Nowok et al., 2016). We used *lavaan* (version 0.6.18; Rosseel, 2012) to estimate confirmatory factor models.

Results

The mean results of the ACO runs are reported in Table 1 for each of the three optimization functions and for scale lengths of 6, 9, 12, 15, 18, and 20 items. The criteria for good model fit, as indicated by the CFI and RMSEA of the CFA model, were met for all $\mathrm{ACO}_{\mathrm{Dim}}$ and $\mathrm{ACO}_{\mathrm{Bal}}$-optimized short scales presented in Table 1. This means that even going as low as six items, short-scales fulfilling these optimization criteria were generated. Regarding McDonald’s ω, values were highest for $\mathrm{ACO}_{\mathrm{Dim}}$ and lowest for $\mathrm{ACO}_{\mathrm{Cat}}$, demonstrating a trade-off between both functions. For $\mathrm{ACO}_{\mathrm{Dim}}$, ω ranged from .86 (6-item version) to .93 (20-item version). Only the full scale showed a higher factor saturation (ω = .97), which is attributed to the length of the scale (99 items). An opposite pattern was found for $r_{\mathrm{polychoric}}$, which was highest for $\mathrm{ACO}_{\mathrm{Cat}}$, with mean values of .90 to .91. Thus, these short scales were even able to compete with the full scale ($r_{\mathrm{polychoric}}$ = .91).

For all indices of interest, the balanced optimization ACO_{Bal} presented a good compromise between ACO_{Cat} and ACO_{Dim}. Figure 1 shows how McDonald's ω and $r_{\text{polychoric}}$ in ACO_{Bal} consistently lie between (or are on par with) the other two optimization functions across all scale lengths tested. Figure 2 additionally shows the density distributions of factor scores across proficiency levels exemplary for the best 12-item short scales of each ACO function. The categorical optimization resulted in an almost equidistant distribution between the mean factor scores, maximizing the factor mean differences between the adjacent proficiency levels. In particular, the mean differences A2/B1 and B1/B2 are larger for the categorical optimization compared to the other optimization forms. While the differences between the three scales appear small, the Figure illustrates that mean differences between proficiency levels could be maximized for a better overall discriminative accuracy by optimizing r_{polychor}, as we expected. Lastly, although we did not specifically include the content coverage of the central language aspects as a criterion in the optimization functions, we observed that a sufficiently broad content coverage was still obtained in many short scales with at least 12 items.

Discussion

Using language proficiency assessment as an example, this study aimed to illustrate the effectiveness of ACO in optimizing ordinal categorical assessment based on several psychometric criteria and to compare three optimization procedures as a proof-of-concept: emphasizing precise dimensional assessment, between-level classification accuracy, or a balance of both. As indicated by the results, short scales could be constructed with ACO that met all four pre-specified criteria. Overall, a very good model fit was easily achieved with all three optimization functions, even if not specifically optimized in ACO_{Cat}. The good model fit values of the different short scales can be attributed to the fact that the item pool of the long scale was already streamlined and well-curated. The same ACO algorithm applied to a more heterogeneous item pool might lead to more significant differences between the original and the optimized scales since it is with large, heterogeneous item pools that metaheuristics such as ACO show their greatest advantages over classical approaches (e.g., Janssen et al., 2017; Leite, 2008).

Most importantly, ACO_{Bal} provided a satisfactory compromise between ACO_{Dim} and ACO_{Cat}. The discrepancy between dimensional and categorical assessment resembles the one between optimizing reliability (i.e., precise dimensional measurement of proficiency levels) and criterion validity (i.e., clear assignment to externally defined proficiency levels). In practice, test developers tend to deal with trade-offs between the reliability and validity of a scale by prioritizing high reliability (Clifton, 2020), because this is easier to achieve either by adding items or by restricting the item pool to highly correlated items. This can lead to what is known as an *attenuation paradox* (Loevinger, 1954), where increases in reliability can be accompanied by decreases in the validity of a scale (Schroeders et al., 2016a). Trade-offs between reliability and validity exist in various forms (Clifton, 2020; Loevinger, 1948). The present study's results showed how ACO can be used to defuse this trade-off by targeting a balance between high criterion validity (categorical assessment) and high reliability (dimensional assessment).

From a psychometric standpoint, it is important to note that we relied on a reflective measurement model for precise, dimensional measurement, while an approach similar to formative modeling was used for the accurate assignment of individuals to proficiency levels (Edwards & Bagozzi, 2000). This choice between reflective and formative assessment is ultimately a question of the measurement intention, that is, explaining the underlying structure of a scale versus predicting an external outcome (see also Shmueli, 2010; Yarkoni & Westfall, 2017). In the present application of ACO, we were interested in compiling short scales that satisfy both objectives. For the sole purpose of categorical assessment, a composite-formative model (Bollen & Diamantopoulos, 2017) may suffice. In this model, a set of observed variables that may even lack conceptual unity forms a weighted continuous (or categorical) composite. Such an approach, which does not rely on specific measurement assumptions (e.g., measurement invariance, high factor saturation), is advantageous when the primary focus is prediction. For instance, in modeling the recidivism risk following release from custody, more or less independent indicators such as prior criminal record, substance abuse, and employment status combine to form a composite score that classifies individuals as either at high or low risk of post-release recidivism (e.g., Goodley et al., 2022). These formative indicators contribute separately to the composite rather than reflecting a common causal latent factor. The composite variable, in turn, allows for a practical, predictive assessment where the goal is to flag individuals at high risk based on multiple factors. In this proof-of-concept study, we decided to demonstrate both psychometric perspectives and applied them to strike a balance between accurate classification and precise dimensional measurement.

A similar ACO procedure balancing categorical and dimensional assessment can be applied to several other areas of psychology. In clinical assessment, instruments could be optimized to assess specific symptom clusters or subtypes of cognitive diseases for adequate treatment provision, such as for attention-deficit/hyperactivity disorder (Galvez-Contreras et al., 2022), while maintaining a robust dimensional assessment of each cluster. ACO could also aid in the construction of efficient short-scales for large-scale educational assessments, where students' scholastic competencies are dimensionally assessed and subsequently assigned to criterion-related competency levels. Metaheuristics could assist the process of finding item sets that increase the contrast between proficiency levels while taking into account other optimization criteria, such as compliance with a pre-specified multidimensional factor structure or the minimization of differential item functioning across demographic groups.

While the present study focuses on ACO, other non-greedy metaheuristics such as Simulated Annealing, Genetic Algorithm, and Tabu Search can be applied with the same model optimization function. In a comparison of these algorithms using Monte Carlo simulations, Raborn et al. (2020) showed that for minor model misspecification, as in the present case, all algorithms produced short scales with good psychometric qualities. For major model misspecification, Simulated Annealing may produce better-fitting short scales. Eventually, the success of an optimization algorithm depends on the specific problem at hand, the quality of the data set, and the hyperparameter settings. None of the mentioned algorithms is universally superior to others when considering the various optimization problems to which it can be applied (Wolpert & Macready, 1997).

To conclude, this study demonstrated the effectiveness and versatility of ACO in constructing short scales for language proficiency assessment, where, for practical reasons, it is often necessary to discretize a continuous latent variable into ordered categories (i.e., language proficiency levels). The ACO algorithm was successful in compiling short scales that, depending on the optimization function, met model fit requirements and achieved high factor saturation (composite reliability) in a reflective measurement framework, a strong correlation with an independent language assessment (criterion validity) in a formative measurement framework, or a trade-off between these two measurement goals.

References

Agresti, A. (2019). *An introduction to categorical data analysis*. Wiley.

Bennette, C., & Vickers, A. (2012). Against quantiles: Categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology*, *12*(1), Article 21. https://doi.org/10.1186/1471-2288-12-21

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods*, *22*(3), 581. https://doi.org/10.1037/met0000056

Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, *45*(1), 41–51. https://doi.org/10.1007/BF00940812

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, *25*(3), 259–270. https://doi.org/10.1037/met0000236

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

DeCoster, J., Gallucci, M., & Iselin, A.-M. R. (2011). Best practices for using median splits, artificial categorization, and their continuous alternatives. *Journal of Experimental Psychopathology*, *2*(2), 197–209. https://doi.org/10.5127/jep.008310

Dorigo, M., & Stützle, T. (2010). Ant Colony Optimization: Overview and recent advances. In M. Gendreau & J.-Y. Potvin (Eds.), *Handbook of Metaheuristics* (pp. 227–263). Springer US. https://doi.org/10.1007/978-1-4419-1665-5

Drezner, Z., & Marcoulides, G. A. (1999). Using simulated annealing for selection in multiple regression analysis. *Multiple Linear Regression Viewpoints*, *25*,(2), 1–4.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2015). Using a genetic algorithm to abbreviate the Psychopathic Personality Inventory–Revised (PPI-R). *Psychological Assessment*, *27*(1), 194–202. https://doi.org/10.1037/pas0000032

Galvez-Contreras, A. Y., Vargas-de la Cruz, I., Beltran-Navarro, B., Gonzalez-Castaneda, R. E., & Gonzalez-Perez, O. (2022). Therapeutic approaches for ADHD by developmental stage and clinical presentation. *International Journal of Environmental Research and Public Health*, *19*(19), Article 12880. https://doi.org/10.3390/ijerph191912880

Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*, *13*(5), 533–549. https://doi.org/10.1016/0305-0548(86)90048-1

Goodley, G., Pearson, D., & Morris, P. (2022). Predictors of recidivism following release from custody: A meta-analysis. *Psychology, Crime & Law*, *28*(7), 703–729. https://doi.org/10.1080/1068316X.2021.1962866

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences*, *66*, 4–15. https://doi.org/10.1016/j.lindif.2017.11.001

Irwin, J. R., & McClelland, G. H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, *40*(3), 366–371. https://doi.org/10.1509/jmkr.40.3.366.19237

Jankowsky, K., Olaru, G., & Schroeders, U. (2020). Compiling measurement invariant short scales in cross–cultural personality assessment using Ant Colony Optimization. *European Journal of Personality*, *34*(3), 470–485. https://doi.org/10.1002/per.2260

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*(6), 409–421. https://doi.org/10.1027/1015-5759/a000299

Kamphuis, J. H., & Noordhof, A. (2009). On categorical diagnoses in DSM-V: Cutting dimensions at useful points? *Psychological Assessment*, *21*(3), 294–301. https://doi.org/10.1037/a0016697

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, *12*(4), 321–344. https://doi.org/10.1080/15305058.2011.643517

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

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

Lenhard, A., Lenhard, W., & Bender, L. (2024). *E-DaF: Einstufungstest für Deutsch als Fremdsprache* [E-DaF - Placement test for German as a foreign language]. Hogrefe.

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of “scale analysis” and factor analysis. *Psychological Bulletin*, *45*(6), 507–529. https://doi.org/10.1037/h0055827

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*(5), 493–504. https://doi.org/10.1037/h0058543

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Marcoulides, K. M., & Falk, C. F. (2018). Model specification searches in structural equation modeling with R. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 484–491. https://doi.org/10.1080/10705511.2017.1409074

Markus, K. A., & Borsboom, D. (2013). Reflective measurement models, behavior domains, and common causes. *New Ideas in Psychology*, *31*(1), 54–64. https://doi.org/10.1016/j.newideapsych.2011.02.008

McClelland, G. H., Lynch, John G., Irwin, J. R., Spiller, S. A., & Fitzsimons, G. J. (2015). Median splits, Type II errors, and false–positive consumer psychology: Don’t fight the power. *Journal of Consumer Psychology*, *25*(4), 679–689. https://doi.org/10.1016/j.jcps.2015.05.006

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus Web Notes*, *4*(5). https://www.statmodel.com/download/webnotes/CatMGLong

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, *74*(11), 1–26. https://doi.org/10.18637/jss.v074.i11

Olaru, G., & Jankowsky, K. (2022). The HEX-ACO-18: Developing an age-invariant HEXACO short scale using ant colony optimization. *Journal of Personality Assessment*, *104*(4), 435–446. https://doi.org/10.1080/00223891.2021.1934480

Olaru, G., Schroeders, U., Hartung, J., & Wilhelm, O. (2019). Ant Colony Optimization and Local Weighted Structural Equation Modeling. A Tutorial on Novel Item and Person Sampling Procedures for Personality Research. *European Journal of Personality*, *33*(3), 400–419. https://doi.org/10.1002/per.2195

Olaru, G., Schroeders, U., Wilhelm, O., & Ostendorf, F. (2019). 'Grandpa, do you like roller coasters?': Identifying age-appropriate personality indicators. *European Journal of Personality*, *33*(3), 264–278. https://doi.org/10.1002/per.2185

Olaru, G., Schroeders, U., Wilhelm, O., & Ostendorf, F. (2018). A confirmatory examination of age-associated personality differences: Deriving age-related measurement-invariant solutions using ant colony optimization. *Journal of Personality*, *86*(6), 1037–1049. https://doi.org/10.1111/jopy.12373

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

Papageorgiou, S. (2016). Aligning language assessments to standards and frameworks. In D. Tsagari & J. Banerjee (Eds.), *Handbook of Second Language Assessment* (pp. 327–340). De Gruyter. https://doi.org/10.1515/9781614513827-022

Raborn, A. W., Leite, W. L., & Marcoulides, K. M. (2020). A comparison of metaheuristic optimization algorithms for scale short-form development. *Educational and Psychological Measurement*, *80*(5), 910–931. https://doi.org/10.1177/0013164420906600

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2.1) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schroeders, U., Scharf, F., & Olaru, G. (2024). Model specification searches in structural equation modeling using Bee Swarm Optimization. *Educational and Psychological Measurement*, *84*(1), 40–61. https://doi.org/10.1177/00131644231160552

Schroeders, U., Wilhelm, O., & Olaru, G. (2016a). Meta-Heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLOS ONE*, *11*(11), Article 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

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Steger, D., Jankowsky, K., Schroeders, U., & Wilhelm, O. (2023). The road to hell is paved with good intentions: How common practices in scale construction hurt validity. *Assessment*, *30*(6), 1811–1824. https://doi.org/10.1177/10731911221124846

Van der Maas, H. L., & Molenaar, P. C. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review*, *99*(3), 395–417. https://doi.org/10.1037/0033-295X.99.3.395

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, 1(1), 67–82. IEEE Transactions on Evolutionary Computation. https://doi.org/10.1109/4235.585893

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* [Doctoral dissertation, University of California].

Zimny, L., Schroeders, U., & Wilhelm, O. (2024). Ant Colony Optimization for parallel test assembly. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-023-02319-7

Table 1

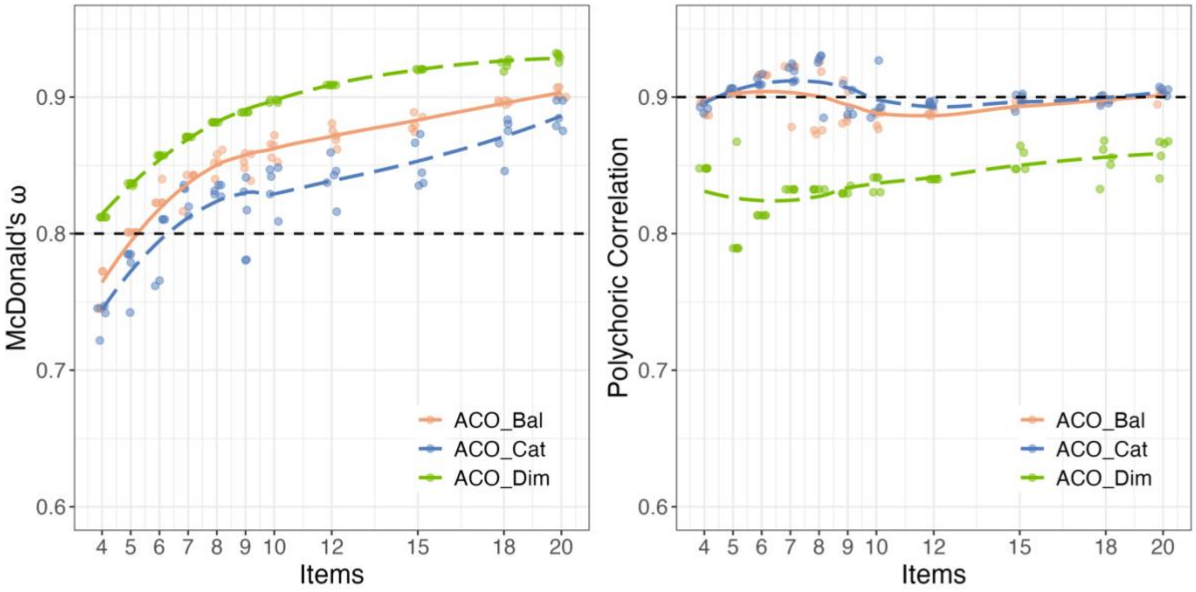
Mean Results by Optimization Procedure and Scale Length

Items	Optimization	φ	CFI	RMSEA	ω	r_{polychor}
6	ACO_Dim	.59 (.00)	1.00 (.00)	.00 (.00)	.86 (.00)	.81 (.00)
6	ACO_Bal	.70 (.00)	1.00 (.00)	.00 (.00)	.83 (.01)	.91 (.01)
6	ACO_Cat	.52 (.08)	1.00 (.00)	.04 (.02)	.79 (.03)	.91 (.00)
9	ACO_Dim	.61 (.00)	1.00 (.00)	.00 (.00)	.89 (.00)	.83 (.00)
9	ACO_Bal	.68 (.02)	1.00 (.00)	.00 (.00)	.85 (.01)	.89 (.01)
9	ACO_Cat	.54 (.07)	.99 (.01)	.03 (.01)	.81 (.03)	.90 (.01)
12	ACO_Dim	.63 (.00)	1.00 (.00)	.00 (.00)	.91 (.00)	.84 (.00)
12	ACO_Bal	.68 (.01)	1.00 (.00)	.00 (.00)	.87 (.01)	.89 (.00)
12	ACO_Cat	.58 (.07)	.99 (.00)	.02 (.01)	.84 (.02)	.90 (.00)
15	ACO_Dim	.65 (.01)	1.00 (.00)	.00 (.00)	.92 (.00)	.85 (.01)
15	ACO_Bal	.70 (.01)	1.00 (.00)	.00 (.00)	.88 (.01)	.90 (.00)
15	ACO_Cat	.55 (.05)	.99 (.00)	.03 (.01)	.85 (.02)	.90 (.01)
18	ACO_Dim	.65 (.02)	1.00 (.00)	.00 (.00)	.92 (.00)	.85 (.01)
18	ACO_Bal	.70 (.01)	1.00 (.00)	.00 (.00)	.90 (.00)	.90 (.00)
18	ACO_Cat	.53 (.03)	.99 (.00)	.03 (.00)	.87 (.02)	.90 (.00)
20	ACO_Dim	.66 (.02)	1.00 (.00)	.00 (.00)	.93 (.00)	.86 (.01)
20	ACO_Bal	.68 (.03)	1.00 (.00)	.00 (.00)	.90 (.00)	.90 (.00)
20	ACO_Cat	.58 (.03)	.99 (.00)	.03 (.00)	.89 (.01)	.90 (.00)
99	Full Scale	-	.99	.02	.97	.91

Note. ACO_{Dim} = dimensional optimization, ACO_{Bal} = balanced optimization, ACO_{Cat} = categorical optimization. Values are the means across the five ACO runs each, with standard deviations in parentheses. φ = overall pheromone value. Rows in bold indicate the minimum item number at which all pre-specified criteria are first met for each optimization procedure.

Figure 1

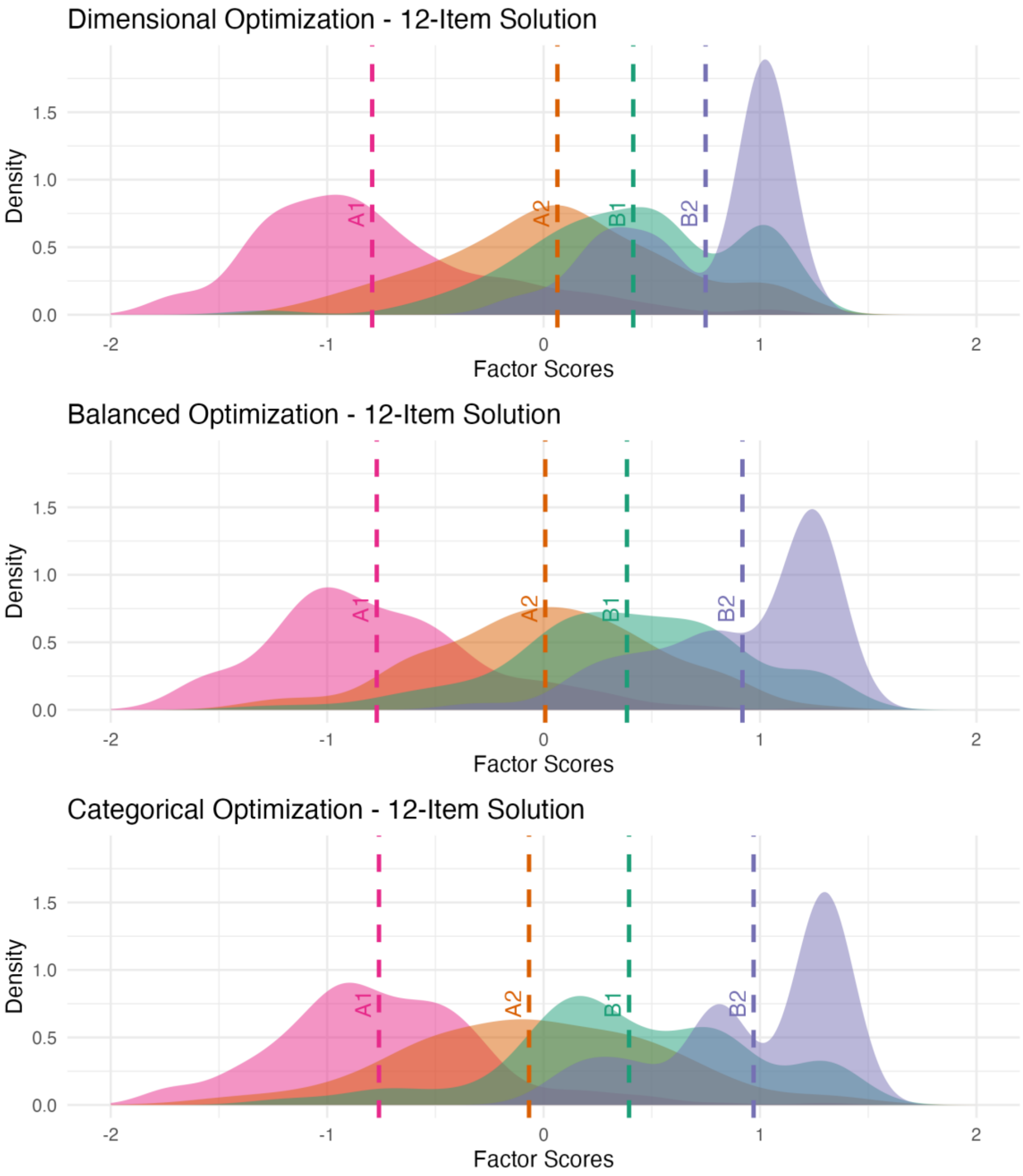
Optimization Criteria as a Function of the Length of the Short Scale



Note. ACO$_{\text{Bal}}$ = balanced optimization, ACO$_{\text{Cat}}$ = categorical optimization, ACO$_{\text{Dim}}$ = dimensional optimization. For each scale length, the dots present the best solutions of each of the 5 ACO runs. Line plots are LOESS-smoothed. The dashed black horizontal line indicates the prespecified cutoff value.

Figure 2

Density Distributions of Language Proficiency Levels Across Factor Scores from the Latent Model for the Best 12-item Scales



Note. N = 581. Factor scores are based on a unidimensional CFA, which was estimated for the best 12-item short scales with dimensional (top panel), balanced (middle panel), and categorical optimization (bottom panel). The mean factor scores of each proficiency level are indicated by the vertical, dashed lines.

Appendix

Functions for the Four Optimization Criteria

$$\varphi\mathrm{CFI} = \frac{1}{1 + \exp\,[75 * (.98 - CFI)]} \tag{1}$$

$$\varphi\mathrm{RMSEA} = 1 - \frac{1}{1 + \exp\,[75 * (.02 - RMSEA)]} \tag{2}$$

$$\varphi\omega = \frac{1}{1 + \exp\,[10 * (.80 - \omega)]} \tag{3}$$

$$\varphi r_{\mathrm{polychor}} = \frac{1}{1 + \exp\,[30 * \left(.90 - r_{\mathrm{polychor}}\right)]} \tag{4}$$

Note. φ = pheromone value. The slope for the logit transformation function was set to allow enough room for optimization, while the threshold value corresponds to the desired minimum values of the respective criteria.

Functions for the Three Optimization Procedures

$$\varphi\mathrm{ACO}_{Dim} = \frac{\frac{1}{2} * (\varphi\mathrm{CFI} + \varphi\mathrm{RMSEA}) + 3 * \varphi\omega)}{4} \tag{5}$$

$$\varphi\mathrm{ACO}_{Cat} = \varphi r_{\mathrm{polychor}} \tag{6}$$

$$\varphi\mathrm{ACO}_{Bal} = \frac{\frac{1}{2} * (\varphi\mathrm{CFI} + \varphi\mathrm{RMSEA}) + 2 * \varphi\omega + 2 * \varphi r_{\mathrm{polychor}})}{5} \tag{7}$$