

## Tilburg University

### The HEX-ACO-18

Olaru, Gabriel; Jankowsky, Kristin

*Published in:*  
Journal of Personality Assessment

*DOI:*  
[10.1080/00223891.2021.1934480](https://doi.org/10.1080/00223891.2021.1934480)

*Publication date:*  
2022

*Document Version*  
Publisher's PDF, also known as Version of record

[Link to publication in Tilburg University Research Portal](#)

*Citation for published version (APA):*  
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>

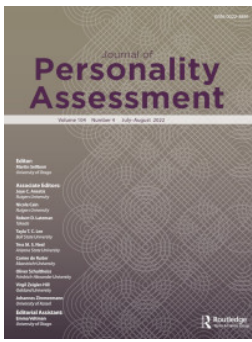
#### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

#### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



# The HEX-ACO-18: Developing an Age-Invariant HEXACO Short Scale Using Ant Colony Optimization

Gabriel Olaru & Kristin Jankowsky

To cite this article: Gabriel Olaru & Kristin Jankowsky (2022) The HEX-ACO-18: Developing an Age-Invariant HEXACO Short Scale Using Ant Colony Optimization, Journal of Personality Assessment, 104:4, 435-446, DOI: [10.1080/00223891.2021.1934480](https://doi.org/10.1080/00223891.2021.1934480)

To link to this article: <https://doi.org/10.1080/00223891.2021.1934480>



© 2021 The Author(s). Published with license by Taylor and Francis Group, LLC



Published online: 17 Jun 2021.



[Submit your article to this journal](#)



Article views: 1401



[View related articles](#)



[View Crossmark data](#)





Citing articles: 1 [View citing articles](#)



This article has been awarded the Centre for Open Science 'Open Materials' badge.

# The HEX-ACO-18: Developing an Age-Invariant HEXACO Short Scale Using Ant Colony Optimization

Gabriel Olaru<sup>1</sup>  and Kristin Jankowsky<sup>2</sup> 

<sup>1</sup>Developmental Psychology, Tilburg University, Tilburg, Netherlands; <sup>2</sup>Psychological Assessment, University of Kassel, Kassel, Germany

## ABSTRACT

In this study, we developed an age-invariant 18-item short form of the HEXACO Personality Inventory for use in developmental personality research. We combined the item selection procedure ant colony optimization (ACO) and the model estimation approach local structural equation modeling (LSEM). ACO is a metaheuristic algorithm that evaluates items based on the quality of the resulting short scale, thus directly optimizing criteria that can only be estimated with combinations of items, such as model fit and measurement invariance. LSEM allows for model estimation and measurement invariance testing across a continuous age variable by weighting participants, rather than splitting the sample into artificial age groups. Using a HEXACO-100 dataset of  $N = 6,419$  participants ranging from 16 to 90 years of age, we selected a short form optimized for model fit, measurement invariance, facet coverage, and balance of item keying. To achieve scalar measurement invariance and brevity, but maintain construct coverage, we selected 18 items to represent three out of four facets from each HEXACO trait domain. The resulting *HEX-ACO-18* short scale showed adequate model fit and scalar measurement invariance across age. Furthermore, the usefulness and versatility of the item and person sampling procedures ACO and LSEM is demonstrated.

## ARTICLE HISTORY

Received 11 February 2021  
Accepted 12 April 2021



Personality traits are robust predictors of a wide variety of relevant life outcomes (Soto, 2019). In many time-constrained contexts, like experience sampling or panel studies, researchers are restricted to very short measures of personality, ranging from one to six items per factor. Popular short measures are the Five or Ten Item Personality Inventory (Gosling et al., 2003), short forms of the Big Five Inventory (e.g., BFI-10; Rammstedt & John, 2007; BFI-SOEP; Hahn et al., 2012; BFI-2-XS; Soto & John, 2017), the mini-IPIP (Donnellan et al., 2006) or the Midlife Development Inventory (MIDI; Lachman & Weaver, 1997). Generally, these scales have been selected from a larger item pool or longer inventory with the goal of maintaining construct coverage or maximizing the internal consistency of the scale. As such, items from the initial item pool were evaluated based on factor loadings, correlations with external outcomes, correlations with the scale scores of the long form, and expert rated construct coverage (e.g. Stanton et al., 2002).

Because of their use in broad and nationally representative panel studies with repeated measurements, these short scales are often used to study change or age-associated differences in personality traits. However, the comparability of the measurement across age (i.e., measurement invariance; MI) was not considered when developing these scales, resulting in potentially non-comparable factor means across

age (e.g., Dong & Dumas, 2020). In addition, nearly all commonly used short scales are Big Five based and neglect the trait domain of Honesty-Humility (Ashton & Lee, 2007). Consequently, developmental findings on this trait domain are scarce, despite its high relevance for many life outcomes not covered by Agreeableness (Ashton et al., 2014). In this study, we seek to develop a HEXACO short scale that fulfills traditional criteria of short scale development, while also providing adequate model fit and MI across the adult lifespan and thus being optimal for use in personality development studies.

## The HEXACO model

The HEXACO model (Ashton et al., 2004) is a six-dimensional alternative to the Big Five (Goldberg, 1990) or Five-Factor personality model (McCrae & Costa, 1996). It has been established and replicated in lexical analyses of several languages (Ashton et al., 2004; Ashton & Lee, 2007). Its main difference to the five-dimensional structures of personality is its sixth Honesty-Humility trait domain, as well as differences in the composition of Agreeableness and Emotionality (i.e., Neuroticism in the five-dimensional models). In the HEXACO model, anger-hostility represents the low pole of the Agreeableness domain, whereas this facet is allocated to the Neuroticism trait domain in the Big Five

**CONTACT** Gabriel Olaru  [golaru@mail.de](mailto:golaru@mail.de)  Developmental Psychology, Tilburg University, Prof. Cobbenhagenlaan 225, DB Tilburg, 5037, Netherlands.

© 2021 The Author(s). Published with license by Taylor and Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

model. Sentimentality is associated with Emotionality (i.e., Neuroticism in the Big Five/FFM) instead of Agreeableness. Inter-individual differences in honesty, fairness, modesty, and greed are only weakly represented in the Agreeableness trait domain of the five-factor models. Whereas the broad NEO-PI-R (Costa & McCrae, 2008)—which is among the longest personality inventories—captures modesty and straightforwardness as facets of Agreeableness, shorter inventories lack a measurement of inter-individual differences in these personality characteristics.

Honesty-Humility is related to proactive cooperation or, conversely, exploitation or cheating (e.g., Heck et al., 2018; Hilbig et al., 2012). The low pole of Honesty-Humility is related to anti-social, egoistical, and psychopathological behavior, encompassed in the dark triad (Muris et al., 2017) or dark factor model (Moshagen et al., 2018). A meta-analytic investigation indicates that Honesty-Humility is more strongly (negatively) related to exploitation (e.g., immoral behavior, short-term mating, lack of cooperation) than Agreeableness (Zettler et al., 2020). In addition, the Honesty-Humility domain is related to political values and orientation (Hilbig & Zettler, 2009), which, in combination with Openness to Experience, can provide a more comprehensive picture of an individual's value system.

Because of the underrepresentation of Honesty-Humility in Big Five-based measures and the lack of short scales capturing this trait domain, relatively little is known about the developmental trajectories thereof. A cross-sectional comparison of the HEXACO mean levels on a large sample of around 100,000 participants showed a linear decrease in Honesty-Humility from 20 to 70 years of age (around one standard deviation), whereas the HEXACO Agreeableness mean-levels differed little across age (Ashton & Lee, 2016). Cross-sectional age patterns of “dark traits” (e.g., deceitfulness, sadism, vindictiveness) showed a linear decline of about one standard deviation from the age of 20 to 50 (Hartung et al., 2021).

However, longitudinal evidence for developmental trends is generally based on short scales used in broad panel studies. To encourage the use of the HEXACO model in contexts with limited item numbers (e.g., panel studies; daily diary studies) and to expand the evidence on developmental trajectories and correlates across the lifespan, we sought to develop a short-scale that provides a high comparability of the HEXACO trait domains across a broad age range.

### **Considerations when developing short scales**

Short scale construction is about defining desired criteria for the scale and identifying the items that maximize these criteria. Traditionally, short scales are developed by selecting the items with the highest Principal Component Analysis main loadings or alpha if item deleted under the assumption that this approach maintains measurement precision and construct coverage of the scale. However, this procedure rewards potentially redundant items with high correlations, resulting in narrow scales that only capture the core aspect of the original constructs. As a result, such a short scale

based on a Conscientiousness item pool may only measure the facet of Diligence, sacrificing validity in favor of internal consistency. Consequently, theoretically assumed correlations with relevant outcomes may not be empirically replicated because the relevant variance has been removed to increase the internal consistency of the scale. Despite also being intended to improve or maintain the unidimensionality of the measures, selecting items with high main loadings does not necessarily improve model fit (Olaru et al., 2015), which is a prevalent issue for broad personality scales (Hopwood & Donnellan, 2010).

Ultra-short scales are the most prevalent measures of personality traits in (longitudinal) panel studies. As such, a large proportion of lifespan personality development research is based on very brief scales. However, in the development of these scales, the comparability of the trait scores across age—and thus an unbiased measurement of personality change—was not adequately addressed. This requirement is also known as the psychometric concept of measurement invariance (MI; Horn & McArdle, 1992). MI refers to the equivalence of the model and parameter estimates (e.g., factor loadings, item intercepts) across groups of people: in the case of aging research, across the age of participants. This constraint ensures that the extracted factors and factor parameters are comparable across age, which is essential when studying change or age-associated differences in the latent traits (Allemand et al., 2007; Brandt et al., 2018; Nye et al., 2016).

A lack of MI can severely bias the findings on developmental trends (e.g., Chen, 2007). It indicates that the items measure different developmental constructs or are affected by age-associated differences that are not captured by the latent factor. To test MI, the measurement model is estimated on different samples (e.g., age groups) and the equivalence of the model parameters is tested across groups. Model parameters are sequentially constrained to equality and the increase in model misfit due to these constraints is evaluated as an indicator of whether MI holds (Cheung & Rensvold, 2002). Generally, a model with no parameter constraints (i.e., configural MI) is compared to a model with equal factor loadings across groups (i.e., metric MI). Subsequently, item intercepts (i.e., scalar MI) and residual variances (i.e., strict MI) are additionally constrained.

The lack of comparability across age is a common issue in personality development research. In an overview of MI testing in personality development studies, Dong and Dumas (2020) reported that 10 out of the 17 (59%) studies achieved scalar MI across age, which is required for the comparison of factor means across age. However, this ratio should be interpreted with caution, as the majority of these studies used very short inventories, aggregated indicators (i.e., item parcels), or only examined a small number of age groups. With an increasing number of items and age groups, achieving MI was less likely (Dong & Dumas, 2020). Many studies also only achieved partial MI, that is, some item intercepts were allowed to vary across age groups. In addition, many personality scales yield insufficient model fit even without invariance constraints (e.g., Hopwood &

Donnellan, 2010), thus technically not even achieving configural measurement invariance. To overcome these issues, model fit and MI need to be considered as item selection criteria when developing scales for personality research and personality development research in particular.

### **The complexity of item selection**

Model fit and MI are scale properties that can only be estimated based on combinations of items. Traditional item selection procedures typically evaluate each item individually, for instance by computing factor loadings, criterion correlations, long form correlation, or construct coverage ratings for each item (e.g., McCrae & Costa, 2004; Donnellan et al., 2006; Rammstedt & John, 2007; Soto & John, 2017). The items are then compared to each other based on the criteria, and the items that excel most with respect to the criteria are then retained for the short form. However, these item-level criteria are only proxies for the scale-level criteria that are of central interest. For instance, items with the lowest main loadings can be eliminated to increase reliability, while items with the highest modification indices can be removed to improve model fit. However, this approach suffers from an unnecessary degree of uncertainty, as items are selected based on their properties in the full model—but not in the final model. When removing items, the characteristics of the remaining items (e.g., factor loadings, modification indices) will change. In addition, when using a large number of selection criteria, item characteristics will not unequivocally support the same items (e.g., item A has the highest main loading but item B the highest outcome correlation). Scale developers must then decide how to balance and weight the various decision rules. However, these decisions are made without knowing what the final short scale properties will be (e.g., are outcome correlations and reliability already high enough but model fit lacking?). When selecting a final subscale, it is thus unclear whether the selected items truly represent the best possible combination of items with respect to the desired scale properties.

A more informed decision on which items should be selected can be reached when the performance of the final short scale with respect to the desired criteria is known. Knowing exactly what model fit the potential resulting short scales will achieve eliminates the uncertainty in the item selection process. However, to do so, each possible combination of items needs to be evaluated. For example, when selecting 15 out of 30 items, this amounts to 155,117,520 possible combinations, for all of which a Confirmatory Factor Analysis would need to be estimated. If MI is also evaluated, the different MI levels also need to be estimated for each item combination, further increasing the computational load. Because estimating such a large number of models is often not feasible due to time and resource constraints, combinatorial optimization algorithms are used to develop psychological short scales (Dörendahl & Greiff, 2020). One such metaheuristic algorithms that has been successfully used as an item selection tool for psychological assessment is ant colony optimization (ACO; e.g., Leite et al., 2008;

Olaru et al., 2015; Olaru, Schroeders, Hartung, et al., 2019). It can be used to select items directly based on the properties of the resulting short scale instead of item-level properties in the initial full model. By applying a heuristic based on the foraging behavior of ants, it is also much more efficient at finding a solution than estimating all possible combinations.

### **Ant colony optimization**

On a general level, ACO is a tool that selects and evaluates combinations of items (for a detailed description see Olaru, Schroeders, Hartung, et al., 2019). Across several iterations, it learns which items provide the best results in term of user-defined optimization criteria (e.g., model fit, reliability). This learning heuristic is mimicking the behavior of ants in search of food. Ants use pheromone trails to communicate shorter routes from the nest to the food source to other ants. In a similar fashion, ACO uses virtual “pheromone” values to identify an optimal item set across several iterations of selecting item combinations and evaluating these on the optimization criterion. For instance, ACO can be used to identify a short scale of six items and evaluate the resulting model fit of a two-factor model based on these items. At the beginning of the search, it will randomly select a number of short scales (the number of item sets to be tested can be set by the user). The model fits of the randomly selected item sets are then compared to one another. The virtual “pheromone” values of the items of the best previous solution are increased, further increasing the likelihood of these items to be drawn in subsequent iterations. Several six-item combinations are selected based on the new selection probabilities and the model fit of the resulting models is again compared to each other. The selection probability is then further increased for the items of the solution with best model fit. This process of selection, evaluation, and increasing of pheromone levels is repeated across several iterations until the desired criteria cannot be further optimized.

### **Local structural equation modeling**

In the current study, we seek to develop a personality short scale that is measurement invariant across age for the use in personality development studies. As such, one central optimization criterion of the scale is its MI across age. MI is generally evaluated with multi-group confirmatory factor analysis (MGCFA), a procedure in which the sample is split into groups and the model is compared across these groups. In the context of personality development research, the sample is divided into age groups. The issue with this approach is that this division represents an artificial categorization of a continuous variable (age) and the resulting groups can be very broad and heterogeneous. Depending on how the sample is split, the group-based findings may differ from the actual underlying age-differences (Hildebrandt et al., 2016; MacCallum et al., 2002). Non-linear developmental trends or potential critical age points are difficult to identify with a low number of broad age groups (Hildebrandt et al., 2016).



In addition, broad age groups result in a loss of information within group differences (MacCallum et al., 2002). Selecting a short scale based on a specific sample split might thus reduce the generalizability of the short scale to applications with a different age distribution.

To address this issue, we used a MI testing procedure that would maintain the continuous nature of age, namely local structural equation modeling (LSEM; Hildebrandt et al., 2009). Instead of splitting the sample into age groups to achieve sufficient sample sizes for the model estimation, LSEM estimates the model at each age point based on sampling weights for the entire sample. Participants with the target age are fully included in the model estimation, whereas younger and older participants are only partially weighted, with decreasing weights based on the difference to the target age (e.g., for the age 30 model, participants with age 30 were weighted by 1.0, 29 and 31-year old participants by 0.9, 28 and 32-year old participants with 0.7, etc.). This weighting function follows a Gaussian distribution around the target age (for an illustration, see Olaru, Schroeders, Hartung, et al., 2019). Because the sampling weights result in overlapping samples for the model estimation at different age points, the model parameters estimated are “smoothed” across the moderator (similar to loess-smoothing).

Using this weighting approach, a sufficient sample size is achieved for the model estimation at each age without having to split the sample into artificial age groups. MI can then be tested similarly to MGCFA by comparing global fit indices (e.g., Comparative Fit Index) between models with parameters constrained to equality. More specifically, a model without additional age-related constraints (i.e., configural MI) is compared to a model with factor loading equivalence across age (i.e., metric MI), which is compared to a model with additionally-constrained item intercepts across age (i.e., scalar MI). As an additional step, residual variances of the items can also be constrained to equality (i.e., strict MI).

### The present study

The goal of the current study is to develop an 18-item HEXACO short scale for the use in personality research, particularly for personality development studies. We chose 18-items, as this would match the length of popular Big Five short scales (e.g., Soto & John, 2017; Hahn et al., 2012), and provide a good tradeoff between construct coverage and developmental homogeneity across facets needed for measurement invariance across age (see, Ashton & Lee, 2016; Olaru et al., 2018). We used the items from the HEXACO-100 and an US American sample of  $N = 6,419$  participants covering a wide age range. We applied ACO to select an 18-item short scale (three items per trait domain) with the goal of maximizing model fit, MI across age in LSEM (i.e., 20 to 70 years of age), construct coverage and item key balance.

## Methods

### Sample

For this study, we reanalyzed data that were collected online at <http://hexaco.org/hexaco-online> (Lee & Ashton, 2020). Participants completed the HEXACO-100 voluntarily and anonymously and were provided feedback after completing the questionnaire. Data were collected from 2014 to 2018. For additional information about the ethics approval, data set and the data cleaning procedure, please see Lee and Ashton (2020). We used a subsample of the data consisting of participants from the USA ( $N = 162,075$ ), as these represented the largest subgroup in the overall sample and would minimize the effect of language and culture on the measurement. We removed participants with a reported age lower than 16 years and higher than 90 years. To ensure that an overrepresentation of younger participants would not skew the weighted samples in LSEM toward younger age and thus affect the item selection procedure, we selected 100 participants matched by gender for each year of age wherever possible (i.e., 16 to 73 years of age). For the age range of 74 to 90 years, we retained all available cases. The resulting sample consisted of 6,419 participants with a mean age of 47.7 years ( $SD = 18.7$ ), of which 3,171 (49%) were female.

Because ACO is an optimization procedure that tries to find the optimal solution for a given sample, it can potentially result in a solution that is *over-fitted* to the specific sample on which it was fit. To ensure that the resulting short scale is generalizable across samples, we cross-validated the scale on a holdout sample. More specifically, we split the original sample into two equally large ( $N = 3,210$  and  $N = 3,209$ ), gender- and age-matched samples. We ran ACO only on the first sample (i.e., *training sample*). We then evaluated the selected short scale by fitting the model on the second sample (i.e., *validation sample*). The weighted training sample sizes in LSEM (i.e., the sum of sample weights ranging from 0 to 1) ranged from  $Nw = 678.8$  for the model at age 20 to  $Nw = 818.8$  at age 70 with a maximum of  $Nw = 939.0$  at age 55. In the validation sample, weighted sample sizes ranged from  $Nw = 676.6$  at age 20 to  $Nw = 834.5$  at age 70 with a maximum of  $Nw = 933.9$  at age 45 (differences between samples resulted from small age differences between training and validation sample:  $M = 47.6/47.8$ ;  $SD = 18.6/18.7$ ).

### Measurement instrument

As the initial item pool, we used the 100-item version of the HEXACO-PI-R (Lee & Ashton, 2018). The HEXACO-PI-R measures six personality trait domains—Agreeableness, Conscientiousness, Emotionality, Extraversion, Honesty-Humility, and Openness—that are further subdivided into four facets, each measured by four items. Of the 25 facet scales, 21 are perfectly balanced with respect to item keying; the remaining four have at least one positively- and one negatively-keyed item. The HEXACO-100 also measures an Altruism facet that is not assigned to any of the aforementioned trait domains. We thus only used the 96 items that

were univocally mapped onto the HEXACO trait domains. Participants indicated their agreement with the different statements on a five-point scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *neutral*, 4 = *agree*, 5 = *strongly agree*). The full questionnaire can be retrieved from the HEXACO homepage (<https://hexaco.org>).

### Statistical analysis

All analysis scripts are available in an OSF repository (<https://osf.io/ayvqt/>). We used a correlated six factor-model with three items per factor. The factors were identified by constraining the first factor loading to 1 and the first item intercept to 0 (i.e., marker method), thus freeing the factor variance and mean. The identification method does not affect the measurement invariance test, as other approaches constrain the factor variance or mean instead, resulting in the same model fit. We used the marker method because it estimates the factor means by default, which are otherwise not automatically freed under scalar measurement invariance constraints in LSEM. We estimated the model with the *cfa*-function in the R-package *lavaan* (Rosseel, 2012). To evaluate MI across age in LSEM, we used the *lsem.estimate*-function in the R package *sirt* (Robitzsch, 2020). In line with the recommendations in the literature (Hildebrandt et al., 2016), we used a bandwidth value of  $h=2$ . We estimated the model from 20 to 70 years of age in steps of five years. Younger and older participants were still included in the model estimation due to the symmetric sample weighting procedure, but the selected end points ensured that enough cases were available on both sides of the weighting function around the target age point. When estimating models at the borders of the age distribution (e.g., below 20 or above 70 years of age), the symmetric weighting function will result in weighted samples that are on average older or younger than the targeted age because of a lack of participants outside these ranges.

By default, LSEM estimates the model sequentially across age (i.e., estimates model parameters for each age point) and provides model fit indices for each age point. To achieve a global model fit estimate for the different MI levels, we used the joint estimation procedure (Robitzsch, 2020). The joint estimation procedure corresponds to a classical MGCFA, in which each weighted age sample is treated as an independent group. Because the weighted samples overlap,  $\chi^2$ -values and degrees of freedom are inflated, while goodness-of-fit indices such as the comparative fit index (CFI) are unaffected. The joint estimation approach also allowed us to constrain model parameters to equality across age for MI testing. We compared the metric MI model (i.e., equal factor loadings across age) to a scalar MI model (i.e., factor loadings and item intercepts constrained to equality across age). We compared the goodness-of-fit index CFI between the nested models to evaluate the increase in misfit due to the parameter constraints. A substantial deviation between subsequently strict models ( $\Delta\text{CFI} > .010$ ; Cheung & Rensvold, 2002) would suggest that the parameters are not measurement invariant across age.

### Optimization criteria

The item selection was set to optimize overall model fit, MI across age, construct coverage, and item key balancing of the resulting 18-item HEXACO short scale. To balance the various optimization criteria equally, we transformed each single criterion to a range from 0 to 1 before computing an average to form the overall evaluation score used by ACO to compare solutions. More details on the optimization criteria are given below (please see OSF for the exact optimization function: <https://osf.io/ayvqt/>).

#### 1. Model fit

To achieve adequate MI and model fit for the short scale, we aimed to maximize the overall model fit of the scalar MI model as this was the most restrictive model we tested. We evaluated overall model fit with a combination of the CFI and the root mean square error of approximation (RMSEA; acceptable/good fit:  $\text{CFI} \geq .90/.95$ ;  $\text{RMSEA} \leq .08/.06$ ; Hu & Bentler, 1999). We thus used the two model fit indices in the ACO optimization function. To ensure that the model fit indicators were comparable despite the different metrics, we logit-transformed each indicator to a range of 0 to 1 (for more details, see Janssen et al., 2017; Olaru, Schroeders, Hartung, et al., 2019).

#### 2. Measurement invariance

To optimize the short scale's MI across age, we tried to reduce the model fit differences between the MI levels. Specifically, we used the CFI difference between the metric and scalar MI level as an optimization criterion with the goal of decreasing it below a threshold of  $\Delta\text{CFI} = .01$ . The CFI difference was also logit transformed. We focused on the scalar measurement invariance level, because this level of measurement invariance was particularly problematic.

#### 3. Construct coverage

To maintain construct coverage, we used two sub-criteria. First, we optimized the proportion of facets covered with a value of 1 for 18 different facets across the 18 items and 0 for a lower coverage/number of facets. We also tried to ensure that the item keying balance was maintained. To do so, solutions with at least four factors including one or two negatively coded items were scored with a 1, and 0 for a lower number of balanced facets.

#### 4. Factor loadings and correlations

In contrast to model fit, MI, and construct coverage, increasing factor loadings was not essential to the scale. Instead, this constraint would have unnecessarily narrowed the construct coverage of the measure. However, to prevent the selection of solutions with good model fit but non-significant or very low factor loadings, the smallest loading of the model was included as an optimization parameter. Specifically, we wanted to ensure that all loadings were at least  $\lambda \geq .30$ .

**Table 1.** Items of the HEX-ACO-18 short scale.

Domain	Facet	Item
H $\omega = .64$ $r_{sj} = .84 (.74)$	Sincerity	I wouldn't pretend to like someone just to get that person to do favors for me. (78)
	Greed-Avoidance	I would like to be seen driving around in a very expensive car. (66 R)
	Modesty	I want people to know that I am an important person of high status. (96 R)
E $\omega = .52$ $r_{sj} = .82 (.70)$	Fearfulness	Even in an emergency I wouldn't feel like panicking. (77 R)
	Dependence	When I suffer from a painful experience, I need someone to make me feel comfortable. (17)
X $\omega = .61$ $r_{sj} = .84 (.75)$	Anxiety	I sometimes can't help worrying about little things. (11)
	Social Self-Esteem	I feel that I am an unpopular person. (52 R)
A $\omega = .58$ $r_{sj} = .84 (.74)$	Social Boldness	I rarely express my opinions in group meetings. (10 R)
	Liveliness	Most people are more upbeat and dynamic than I generally am. (94 R)
C $\omega = .61$ $r_{sj} = .78 (.63)$	Forgiveness	I rarely hold a grudge, even against people who have badly wronged me. (3)
	Gentleness	I generally accept people's faults without complaining about them. (33)
O $\omega = .52$ $r_{sj} = .80 (.67)$	Patience	I find it hard to keep my temper when people insult me. (93 R)
	Diligence	Often when I set a goal, I end up quitting without having reached it. (56 R)
	Prudence	I make a lot of mistakes because I don't think before I act. (44 R)
	Organization	When working, I sometimes have difficulties due to being disorganized. (74 R)
	Unconventionality	I think that paying attention to radical ideas is a waste of time. (19 R)
	Aesthetic Appreciation	If I had the opportunity, I would like to attend a classical music concert. (49)
	Creativity	I would enjoy creating a work of art, such as a novel, a song, or a painting. (37)

Note.  $\omega$  = Reliability (factor saturation) McDonald's omega;  $r_{sj}$  = correlation between the scale scores of the short and long form (part-whole corrected). The original HEXACO-100 item number is given in parentheses, R indicates that the item is reverse-keyed.

Another criterion we deemed necessary for a useful short scale is that the factors should not highly overlap, as is often the case for self-report personality measurement models (e.g., Park et al., 2020). High correlations between the trait domains may indicate strong response styles or artifacts, such as socially desirable responding or self-evaluation tendencies (Leising et al., 2020). Selecting items that would decrease trait domain correlations might thus reduce the effect of these sources of variance on the measurement of the personality traits. The largest absolute factor correlations were logit-transformed with a turning point of  $r = .55$ : the objective was to minimize these values. This cutoff may seem high, but the personality trait domain correlations are typically high when estimated with Confirmatory Factor Analysis (Park et al., 2020).

### Ant colony optimization parameters

We estimated 120 models per iteration. After each iteration, pheromone values for the items of the best solution found in the iteration were increased by the evaluation score (i.e., the average across all optimization criteria ranging from 0 to 1). The search was aborted if no improvement to the overall best solution could be found after 90 iterations. As ACO is a probabilistic procedure that may yield a different solution with each run, we started the item selection 20 times with different random number generator seeds and used the overall best solution out of the 20 runs (based on the training sample).

## Results

### Construct coverage and item key balance

The final item set selected by ACO (HEX-ACO-18) is presented in Table 1. Each item represents a different facet from the original HEXACO model. The facets Fairness (H), Sentimentality (E), Sociability (X), Flexibility (A), Perfectionism (C), and Inquisitiveness (O) were not retained in the final scale. Item key balance was mainly maintained,

with all factors except for Conscientiousness and Extraversion containing at least one positive and one negatively coded item. Overall, there were seven positively keyed and eleven negatively keyed items in the final short scale. The average correlation between the scale scores of the short and long form was  $r = .82$  (see Table 1).

### Factor loadings and correlations

The final six factor-model estimated on the validation sample is presented in Figure 1. As can be seen, all factor loadings were at least  $\lambda \geq .30$ . The average factor saturation McDonald's omega was  $\omega = .58$  for the short scale (see Table 1). In the long form, the average factor saturation was  $\omega = .84$  (ranging from Conscientiousness  $\omega = .81$  to Honesty-Humility  $\omega = .87$ ), which would correspond to  $\omega = .50$  for a three item scale. The average absolute correlation between the factors was  $r = .24$  (long form mean absolute  $r = .23$ ), and similar to the factor analytic correlations in the long form (for a comparison see OSF Figure 1; <https://osf.io/ayvqt/>). The factor analytic correlations between Extraversion, Emotionality and Agreeableness were highest (see OSF Table 1 and 2 for scale score correlations), and we were not able to reduce them in the process of item selection. The most notable deviations between the short and long form were a smaller Honesty-Humility—Agreeableness correlation ( $r_{\text{short}} = .39$  vs.  $r_{\text{long}} = .49$ ), and a stronger negative Emotionality—Agreeableness correlation ( $r_{\text{short}} = -.52$  vs.  $r_{\text{long}} = -.37$ ). These differences can be attributed to the omission of Sentimentality, Flexibility and Fairness, which correlated positively with the other trait domains (see OSF Table 3).

### Model fit and measurement invariance

Model fit for the full 18-item model was acceptable in both the training and validation sample (see Table 2). Model fit decreased when tested out of the original training sample but was still adequate. Constraining factor loadings and item



intercepts to equality across age did not deteriorate model fit beyond the common cutoffs. Scalar MI could thus be achieved in both the training and validation sample (see Table 2), suggesting that the factor means based on the selected scale were comparable across a broad age range.

**The power of automated combinatorial item selection**

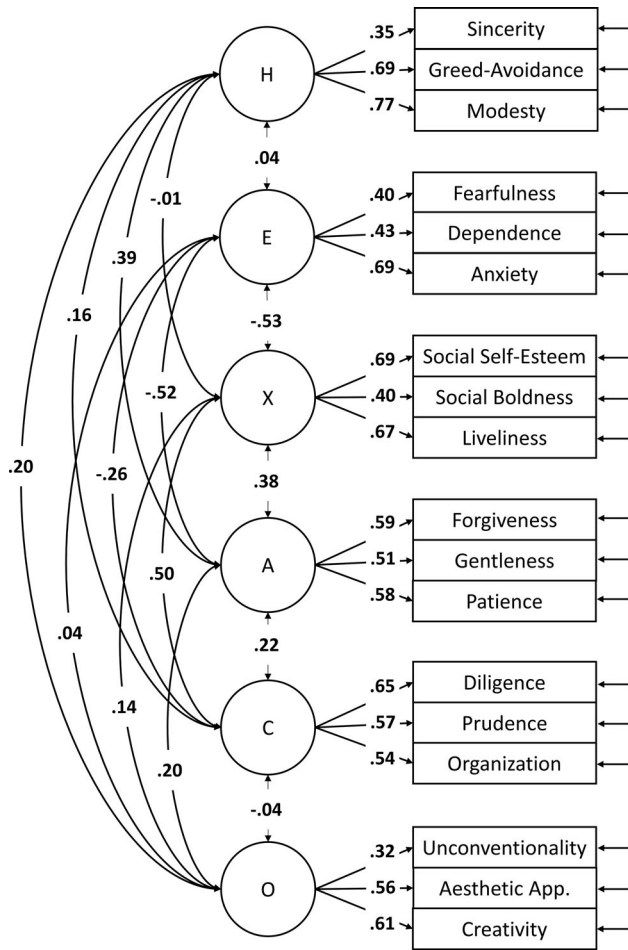
To demonstrate the effectiveness of ACO as an item selection technique, we compared the achieved criteria to that of

a random selection of items. More specifically, we generated 1,000 item sets that would adhere to the same constraints on facet coverage and item balance as our selected short scale (i.e., 18 facets and item key balance on at least four trait domains). We then estimated their corresponding models. The distributions of CFI, RMSEA, and ΔCFI between metric and scalar measurement invariance levels are presented in Figure 2, alongside the achieved values of our item selection. As illustrated, ACO optimized all criteria beyond the 99.9<sup>th</sup> percentile of the distributions, in both the training and validation sample. In contrast to the RMSEA, which was acceptable for all randomly selected solutions, the CFI and ΔCFI seemed to be particularly problematic. The CFI is generally poor for broad personality models, as the relatively low factor loadings and large number of small cross-loadings or residual correlations may decrease the difference between the tested and null or baseline model (see e.g., Moshagen & Auerwald, 2018). Measurement invariance across age is generally difficult to achieve because of the heterogeneity of the (facet-)items and the age-associated patterns thereof.

**Discussion**

In this study, we developed a HEXACO short scale for use in developmental research and personality research in general. We selected 18 items from the HEXACO-100 to create a short scale that would achieve sufficient model fit and MI across age, while maintaining construct coverage and item key balance. Achieving good model fit with broad personality scales covering several trait domains is a very difficult task (Hopwood & Donnellan, 2010; see also Figure 2) that requires state-of-the-art item selection tools such as ACO. Because we optimized the scale directly within a six-factor model instead of separate one-factor models, we also ensured that residual correlations and cross-loadings would be minimized and that the entire scaled could be modeled at once. By also reducing the trait domain correlations, we tried to ensure that the resulting scales would be as unidimensional and “pure” as possible, without resorting to approaches that would narrow down the construct coverage.

Because we optimized MI across age, the HEX-ACO-18 is not only suitable for studying normative, structural, and divergent age patterns, but also for how outcome correlations vary across age or life stages. However, the scale

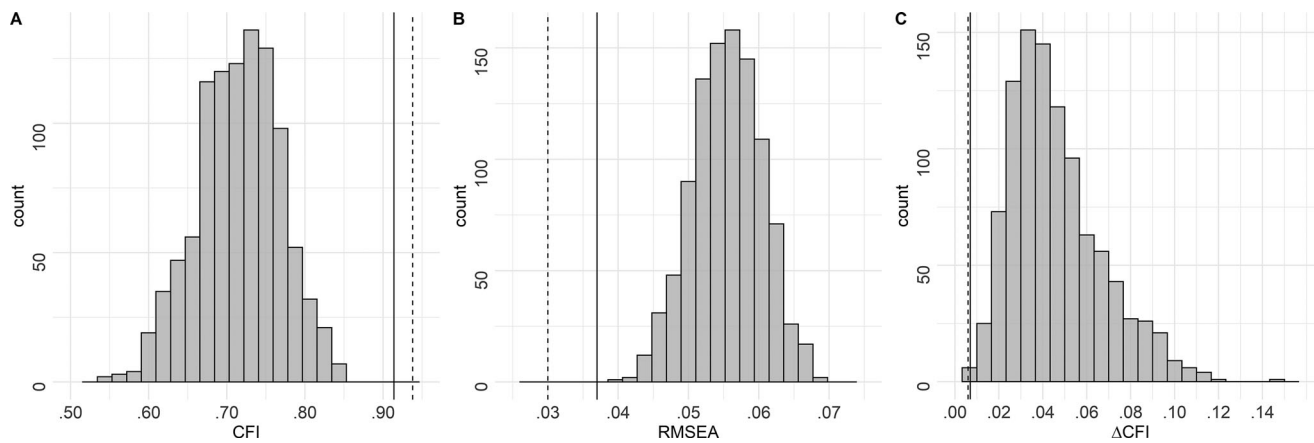


**Figure 1.** Model parameters of the HEX-ACO. Note. Model parameters were estimated based on the scalar measurement invariant model in the full validation sample (N=3,209). For model fit indices see Table 2.

**Table 2.** Model fit and measurement invariance across age.

CFA	df	Training			Validation				
		$\chi^2$	CFI	RMSEA	SRMR	$\chi^2$	CFI	RMSEA	SRMR
Full sample	120	308.360	.931	.036	.031	823.793	.908	.043	.036
LSEM			CFI	RMSEA	SRMR		CFI	RMSEA	SRMR
Configural			.942	.032	.035		.920	.039	.039
Metric			.944	.030	.036		.921	.037	.040
Scalar			<b>.938</b>	<b>.030</b>	<b>.038</b>		<b>.914</b>	<b>.037</b>	<b>.042</b>
Strict			.909	.035	.042		.887	.040	.046

Note. CFA = Confirmatory factor analysis; Full sample = Model estimated on the full training/validation sample without age-moderation; LSEM = Local structural equation modeling (used for measurement invariance testing); Configural = No equality constraints across samples/moderators; metric = constrained factor loadings across age; scalar = additionally constrained factor item intercepts across age; strict = additionally constrained item residuals across age; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual. Degrees of freedom and  $\chi^2$ -values for LSEM are inflated due to treating the weighted samples as independent and are thus not reported. Achieved measurement invariance levels are marked in bold.



**Figure 2.** Model fit and measurement invariance distribution of randomly selected 18-item models.

*Note.* Distributions of CFI, RMSEA and  $\Delta$ CFI of 1,000 randomly selected scalar measurement models. The dashed and solid vertical lines indicate the achieved criterion for the HEX-ACO solution in the training sample and validation sample, respectively.

development approach used in this study also ensures that the scale is useful for personality research in general. Personality scales are often developed on restricted age samples (e.g., students, middle-aged adults) and thus the applicability of the selected items outside this age range may not be given (e.g., Olaru, Schroeders, Wilhelm, et al., 2019). Because we selected items in such a way that the psychometric properties were stable across the adult lifespan, the scale can be used in student samples just as well as in studies on old adulthood.

Defining central criteria, specifying their relative importance and desired or acceptable cutoff values is an essential step in scale construction. No item set will excel on all properties, and some selection criteria may be opposing (e.g., construct heterogeneity and measurement invariance). For the current goal of developing an unbiased assessment of personality change or age-associated differences, we deemed model fit and measurement invariance across age most relevant. As such, we prioritized these characteristics over others. Consequently, we could not maintain full construct coverage and perfect item key balance. For instance, the Extraversion and Conscientiousness scales are only represented by reverse-keyed items, and as such may be more affected by acquiescent or extreme response styles than the other trait scales. The facets Fairness (H), Sentimentality (E), Sociability (X), Flexibility (A), Perfectionism (C), and Inquisitiveness (O) were not retained for the HEX-ACO-18. The omission of these facets has shifted the construct coverage of some of the trait domains. Most notably, the omission of Inquisitiveness from the Openness factor resulted in a stronger cultural emphasis of the scale and a reduced its focus on intellect. Similarly, the exclusion of Fairness from the Honesty-Humility domain resulted in a stronger representation of the Humility aspect of the trait domain, more specifically through the inclusion of two status-seeking items.

The excluded facets diverged most from the trait domain age patterns in a study on cross-sectional age differences in the full HEXACO (Ashton & Lee, 2016). For scalar MI across age at the trait domain level, the age patterns of the facets must be similar (e.g., Olaru et al., 2018). In line with

this, ACO discarded the facets that deviated most strongly from this pattern. Conceptually, the trait domain scale was cleaned in such a way that a cohesive developmental construct was measured, at the cost of within-age specificity. As such, this approach is in stark contrast to studies on age-associated differences in the nuances (i.e., items) of personality (Möttus et al., 2019; Möttus & Rozgonjuk 2021), as we eliminated age-associated differences on the item and facet level that were not accounted for by the trait domain level (for a discussion on measurement invariance and formative models of personality see, Achaa-Amankwaa et al., 2021). Formative models can be particularly interesting in the context of broad and heterogenous short scales to maximize the scale-outcome associations (Myszkowski et al., 2019). In this case, the explained variance by the six trait domain scores could be compared to the 18 facet-items to judge how much of the outcome related variance is specific to the facets—and whether a longer measure is needed to investigate the associations in more detail.

Short scales are often criticized for their low reliability, but this view is often based on the traditional focus on internal consistency of scales—which provides an under-estimation of the actual item reliabilities (Sijtsma, 2009). For a three-item scale, achieving a McDonald's omega or Cronbach's alpha of  $\omega/\alpha \geq .70$  would require the average factor loading to be at least  $\lambda = .66$ , or the average inter-item correlation to be  $r = .43$ , respectively. These correlations are quite high for heterogeneous trait domain measures and would narrow the construct coverage. A more adequate representation of reliability would be based on the test-retest correlations of the scale. Recent studies have shown that single personality items have quite high retest correlations, with an average of  $r = .65$  (Möttus et al., 2019) over the course of one or two-weeks. For a scale of three items, the scale level retest reliability would amount to  $r = .85$ .

### Short scale construction as a combinatorial problem

Metaheuristic optimization algorithms such as ACO are powerful scale development tools because they are able to optimize combinatorial scale properties (e.g., model fit)

directly on the resulting short scale. Traditionally, items are selected based on item level properties (e.g., factor loading) of the initial item pool (e.g., the full scale). This approach relies on the assumption that the item level criteria are directly related to the desired scale criteria and that the initial item properties do not change after items have been removed. In addition, the various item selection criteria generally do not unequivocally support the same items: it is up to the researcher to decide which items to choose for the final version. However, as Figure 2 showed, the room for errors or uncertainty in the item selection procedure is small to non-existent when optimizing model fit and measurement invariance of broad personality scales (see also, Jankowsky et al., 2020). Because the most optimal approach—estimating each possible combination of items—is too computationally demanding, we used ACO as a heuristic item selection procedure to reduce the number of models that had to be estimated.

This study has also shown the usefulness of combinatorial item-selection for the field of psychological assessment, but in particular personality assessment, which suffers from the issues of a lack of model fit (e.g., Hopwood & Donnellan, 2010) and MI (Dong & Dumas; 2020). ACO or similar metaheuristic procedures (e.g., genetic algorithm; Yarkoni, 2010) have been used in several scale development contexts to improve (among others) model fit and reliability (e.g., Kerber et al., 2020; Leite et al., 2008; Janssen et al., 2017). They have also been used in several studies to improve the MI of scales in a multi-group context, for instance to improve the gender-fairness of knowledge tests (Schroeders et al., 2016) or the comparability between personality scales across different languages or cultures (Jankowsky et al., 2020; Olaru & Danner, 2021). Typically, scales are first developed based on English-speaking samples and then translated and applied in other cultural contexts. However, this approach may cause measurement issues due to cultural differences (e.g., the Excitement-Seeking item “I like roller coasters” in the Philippines; Church et al., 2011). Similar issues can be found when scales are developed for young or middle-aged adults but applied to samples of older age afterwards (e.g., “I like roller coasters” for older participants; Olaru, Schroeders, Wilhelm, et al., 2019).

So far, metaheuristic optimization algorithms have been mostly used in the context of short scale development based on established measurement instruments, but rarely to develop full scales (for an exception, see, Moshagen et al., 2018). In the context of short scale construction, the factor structure is generally given by the long form and items are selected within this model. When developing a full scale, items can also be based on a theoretically-assumed structure and selected with ACO to fit this structure well. Because ACO only selects the items, any type of structure can be defined by the user (e.g., higher-order; Olaru et al., 2018; bi-factor; Jankowsky et al., 2020). For a more bottom-up development approach (e.g., Condon et al., 2020), a large item pool can be generated or created based on existing measures (e.g., Condon, 2018; Yarkoni, 2010). After an initial structuring or item clustering step, items can then be selected with ACO to fit the derived factor structure best (see e.g., Wendt et al., 2021).

### **Continuous age moderation with local structural equation modeling**

We used LSEM in this study to moderate the measurement model across age and test for MI. In personality development research, age patterns in the personality traits are investigated based on scale scores or in MGCFA if latent variable modeling is used. The issue with both approaches is that age groups have to be created before scores or factor means are compared across age. This step is always somewhat arbitrary as age differences within a group can be larger than between individuals in different groups (e.g., at the cutoffs). The way in which age groups are formed will affect the overall patterns found (Hildebrandt et al., 2016; MacCallum et al., 2002). Even if the overall sample size allows for narrow age groups, individuals within the group may be in different life stages (e.g., transition from education into work) despite similar chronological age. The discrepancy between chronological age and other more subjective age concepts also applies to LSEM, but the weighting approach can somewhat reduce this difference. With a median age of marriage around 30 years, an age group approach might split the sample into an early and late marriage group, whereas the LSEM model at age 30 will include all participants around that age in the model estimation.

We used LSEM in a cross-sectional context, but its true potential lies in the combination of LSEM and longitudinal models (Olaru et al., 2020; Wagner et al., 2019). Longitudinal models of personality can be moderated across continuous age to investigate how personality change across time varies as a function of participants' age. Because age is incorporated as a within- and between-subject variable in these models, researchers can distinguish between cross-sectional and longitudinal age differences and check if these align. As each parameter of the model is moderated, this approach also allows for the simultaneous examination of normative (i.e., mean-level), differential (i.e., rank-order stability), structural (i.e., correlations) and divergent (i.e., variances) personality change across time and age.

Cross-sectional and longitudinal studies are complementing approaches of studying age-associated differences in personality. Longitudinal approaches allow for the examination of change by studying the same individuals over repeated measurement occasions, whereas cross-sectional studies are focused on age-associated differences between participants. Cross-sectional studies generally cover broader age ranges and item sets, but may be affected by cohort differences. Longitudinal findings may be affected by measurement occasion effects or repeated scale administration (e.g., regression to the mean, practice effects). In this study we optimized MI across age in a cross-sectional context. An optimization of both cross-sectional and longitudinal comparability with longitudinal LSEM would have been preferable. However, longitudinal datasets generally only cover narrow age ranges or use shortened inventories that cannot be used in the context of item selection. In addition, longitudinal administrations of longer personality inventories generally only cover short time spans used to evaluate re-test reliability or a small number of repeated measures, which would not allow us to rule out measurement occasion specific effects.

## Limitations of the present study

Although the short scale in this study was derived from a well-established measure of the HEXACO model and a broad age sample of over 6,000 participants, it still suffers from some limitations. First, the sample used is an online convenience sample. As such, the quality of the data might have been compromised. On the other hand, as participants only received feedback on their trait levels and no other reward, the incentive for faking or careless responding should have been rather small. Second, because the sample was a convenience sample, it might represent a more educated proportion of the population with an interest in personality. Third, even though we were able to select from 100 items, a larger item pool (e.g., HEXACO-200) would have been preferable to increase the quality of the final short scale. However, because age-coverage was of central concern in this study, we decided to use the much broader sample collected for the 100-item version. And finally, to achieve a sufficiently low item number, as well as adequate model fit and MI across age, we could only retain three out of four facets per trait domain. The short scale thus does not provide full HEXACO facet coverage.

## Conclusion

In this study, we developed an 18-item HEXACO short scale for use in personality research, particularly in developmental studies. We used ACO to select items from the HEXACO-100 that would optimize model fit, measurement invariance across age, and construct coverage of the resulting short scale. We used LSEM to estimate age effects on the measurement of the HEXACO trait domains and maintain the continuous nature of age. The HEX-ACO-18 covers 18 of the original HEXACO facets and yielded adequate model fit and measurement invariance across a broad age range of 20 to 70 years of age. In addition to developing a short scale, we also demonstrated the purposefulness and versatility of metaheuristic item selection procedures in the context of psychological assessment. We also showed the usefulness of LSEM for developmental research.

## Acknowledgments

We would like to thank Kibeom Lee and Michael C. Ashton for sharing the data used in this study and providing comments on an earlier version of the manuscript.

## Open Scholarship



This article has earned the [Center for Open Science](https://osf.io/ayvqt/) badge for Open Materials through Open Practices Disclosure. The materials are openly accessible at <https://osf.io/ayvqt/>.

## Declaration of interest statement

We have no conflicts of interest to disclose.

## ORCID

Gabriel Olaru  <http://orcid.org/0000-0002-7430-7350>  
 Kristin Jankowsky  <http://orcid.org/0000-0002-4847-0760>

## Data availability statement

The data was obtained upon request from Kibeom Lee and Michael C. Ashton (Lee & Ashton, 2020).

## References

- Achaa-Amankwaa, P., Olaru, G., & Schroeders, U. (2021). Coffee or tea? Examining cross-cultural differences in personality nuances across former colonies of the British Empire. *European Journal of Personality, 35*(3), 383–397. <https://doi.org/10.1177/0890207020962327>
- Allemand, M., Zimprich, D., & Hertzog, C. (2007). Cross-sectional age differences and longitudinal age changes of personality in middle adulthood and old age. *Journal of Personality, 75*(2), 323–358. <https://doi.org/10.1111/j.1467-6494.2006.00441.x>
- Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc, 11*(2), 150–166. <https://doi.org/10.1177/1088868306294907>
- Ashton, M. C., & Lee, K. (2016). Age trends in HEXACO-PI-R self-reports. *Journal of Research in Personality, 64*, 102–111. <https://doi.org/10.1016/j.jrp.2016.08.008>
- Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc, 18*(2), 139–152. <https://doi.org/10.1177/1088868314523838>
- Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., Boies, K., & De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86*(2), 356–366. <https://doi.org/10.1037/0022-3514.86.2.356>
- Brandt, N. D., Becker, M., Tetzner, J., Brunner, M., Kuhl, P., & Maaz, K. (2018). Personality across the lifespan. *European Journal of Psychological Assessment, 36*(1), 162–173. <https://doi.org/10.1027/1015-5759/a000490>
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. <https://doi.org/10.1080/10705510701301834>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255. [https://doi.org/10.1207/S15328007SEM0902\\_5](https://doi.org/10.1207/S15328007SEM0902_5)
- Church, A. T., Alvarez, J. M., Mai, N. T., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the revised NEO personality inventory. *Journal of Personality and Social Psychology, 101*(5), 1068–1089. <https://doi.org/10.1037/a0025290>
- Condon, D. M. (2018, January 10). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. <https://doi.org/10.31234/osf.io/sc4p9>
- Condon, D. M., Wood, D., Möttus, R., Booth, T., Costantini, G., Greiff, S., ... Zimmermann, J. (2020, December). Bottom up construction of a personality taxonomy. <https://doi.org/10.31234/osf.io/u2n7s>
- Costa, P. T., Jr, & McCrae, R. R. (2008). *The revised NEO personality inventory (NEO-PI-R)*. Sage Publications, Inc.
- Dong, Y., & Dumas, D. (2020). Are personality measures valid for different populations? A systematic review of measurement invariance



- across cultures, gender, and age. *Personality and Individual Differences*, 160, 109956. <https://doi.org/10.1016/j.paid.2020.109956>
- Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, 18(2), 192–203. <https://doi.org/10.1037/1040-3590.18.2.192>
- Dörendahl, J., & Greiff, S. (2020). Are the machines taking over? Benefits and challenges of using algorithms in (short) scale construction [Editorial]. *European Journal of Psychological Assessment*, 36(2), 217–219. <https://doi.org/10.1027/1015-5759/a000597>
- Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6), 1216–1229. <https://doi.org/10.1037/0022-3514.59.6.1216>
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B. Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6), 504–528. [https://doi.org/10.1016/S0092-6566\(03\)00046-1](https://doi.org/10.1016/S0092-6566(03)00046-1)
- Hahn, E., Gottschling, J., & Spinath, F. M. (2012). Short measurements of personality—Validity and reliability of the GSOEP Big Five Inventory (BFI-S). *Journal of Research in Personality*, 46(3), 355–359. <https://doi.org/10.1016/j.jrp.2012.03.008>
- Hartung, J., Bader, M., Moshagen, M., & Wilhelm, O. (2021). Age and gender differences in socially aversive ("dark") personality traits. *European Journal of Personality*, Advance online publication. <https://doi.org/10.1177/0890207020988435>
- Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making*, 13(4), 356.
- Hilbig, B. E., & Zettler, I. (2009). Pillars of cooperation: Honesty–Humility, social value orientations, and economic behavior. *Journal of Research in Personality*, 43(3), 516–519. <https://doi.org/10.1016/j.jrp.2009.01.003>
- Hilbig, B. E., Zettler, I., & Heydasch, T. (2012). Personality, punishment and public goods: Strategic shifts towards cooperation as a matter of dispositional honesty–humility. *European Journal of Personality*, 26(3), 245–254. <https://doi.org/10.1002/per.830>
- Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., & Wilhelm, O. (2016). Exploring factor model parameters across continuous variables with local structural equation models. *Multivariate Behavioral Research*, 51(2-3), 257–258. <https://doi.org/10.1080/00273171.2016.1142856>
- Hildebrandt, A., Wilhelm, O., & Robitzsch, A. (2009). Complementary and competing factor analytic approaches for the investigation of measurement invariance. *Review of Psychology*, 16(2), 87–102.
- Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, 14(3), 332–346. <https://doi.org/10.1177/1088868310361240>
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18(3-4), 117–144. <https://doi.org/10.1080/03610739208253916>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- 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>
- Kerber, A., Schultze, M., Müller, S., Rühling, R. M., Wright, A. G., Spitzer, C., ... Zimmermann, J. (2020). Development of a short and ICD-11 compatible measure for DSM-5 maladaptive personality traits using ant colony optimization algorithms. Advance online publication. *Assessment*. <https://doi.org/10.1177/1073191120971848>
- Lachman, M. E., & Weaver, S. L. (1997). *The midlife development inventory (MIDI) personality scales: Scale construction and scoring*. Brandeis University, pp. 1–9.
- Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment*, 25(5), 543–556. <https://doi.org/10.1177/1073191116659134>
- Lee, K., & Ashton, M. C. (2020). Sex differences in HEXACO personality characteristics across countries and ethnicities. *Journal of Personality*, 88(6), 1075–1090. <https://doi.org/10.1111/jopy.12551>
- Leising, D., Vogel, D., Waller, V., & Zimmermann, J. (2020). Correlations between person-descriptive items are predictable from the product of their mid-point-centered social desirability values. *European Journal of Personality*. Advance online publication. <https://doi.org/10.1177/0890207020962331>
- 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>
- 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>
- McCrae, R. R., & Costa, P. T. Jr. (2004). A contemplated revision of the NEO five-factor inventory. *Personality and Individual Differences*, 36(3), 587–596. [https://doi.org/10.1016/S0191-8869\(03\)00118-1](https://doi.org/10.1016/S0191-8869(03)00118-1)
- McCrae, R. R., & Costa, P. T. Jr. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J.S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 51–87). Guilford.
- Moshagen, M., & Auerswald, M. (2018). On congruence and incongruence of measures of fit in structural equation modeling. *Psychological Methods*, 23(2), 318–336. <https://doi.org/10.1037/met0000122>
- Moshagen, M., Hilbig, B. E., & Zettler, I. (2018). The dark core of personality. *Psychological Review*, 125(5), 656–688. <https://doi.org/10.1037/rev0000111>
- Möttus, R., & Rozgonjuk, D. (2021). Development is in the details: Age differences in the Big Five domains, facets, and nuances. *Journal of Personality and Social Psychology*, 120(4), 1035–1048. <https://doi.org/10.1037/pspp0000276>
- Möttus, R., Sinick, J., Terracciano, A., Hřebíčková, M., Kandler, C., Ando, J., Mortensen, E. L., Colodro-Conde, L., & Jang, K. L. (2019). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability and utility of personality nuances. *Journal of Personality and Social Psychology*, 117(4), e35–e50. <https://doi.org/10.1037/pspp0000202>
- Muris, P., Merckelbach, H., Otgaar, H., & Meijer, E. (2017). The mal-eval side of human nature: A meta-analysis and critical review of the literature on the dark triad (narcissism, Machiavellianism, and psychopathy). *Perspectives on Psychological Science*, 12(2), 183–204. <https://doi.org/10.1177/1745691616666070>
- Myszkowski, N., Storme, M., & Tavani, J. L. (2019). Are reflective models appropriate for very short scales? Proofs of concept of formative models using the ten-item personality inventory. *Journal of Personality*, 87(2), 363–372. <https://doi.org/10.1111/jopy.12395>
- Nye, C. D., Allemand, M., Gosling, S. D., Potter, J., & Roberts, B. W. (2016). Personality trait differences between young and middle-aged adults: Measurement artifacts or actual trends? *Journal of Personality*, 84(4), 473–492. <https://doi.org/10.1111/jopy.12173>
- Olaru, G., & Danner, D. (2021). Developing cross-cultural short scales using ant colony optimization. *Assessment*, 28(1), 199–210. <https://doi.org/10.1177/1073191120918026>
- 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. (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., 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., 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>
- Olaru, G., Robitzsch, A., Hildebrandt, A., Schroeders, U. (2020, April 24). Local structural equation modeling for longitudinal data. <https://doi.org/10.31234/osf.io/q79c5>
- Park, H. H., Wiernik, B. M., Oh, I. S., Gonzalez-Mulé, E., Ones, D. S., & Lee, Y. (2020). Meta-analytic five-factor model personality inter-correlations: Eeny, meeny, miney, moe, how, which, why, and where to go. *Journal of Applied Psychology*, 105(12), 1490–1529. <https://doi.org/10.1037/apl0000476>
- Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *Journal of Research in Personality*, 41(1), 203–212. <https://doi.org/10.1016/j.jrp.2006.02.001>
- Robitzsch, A. (2020). SIRT: Supplementary item response theory models. R package version 3.9-4. <https://CRAN.R-project.org/package=sirt>
- 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., Wilhelm, O., & Olaru, G. (2016). 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>
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107. <https://doi.org/10.1007/s11336-008-9101-0>
- Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, 30(5), 711–727. <https://doi.org/10.1177/0956797619831612>
- Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the big five inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, 68, 69–81. <https://doi.org/10.1016/j.jrp.2017.02.004>
- Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, 55(1), 167–194. <https://doi.org/10.1111/j.1744-6570.2002.tb00108.x>
- Wagner, J., Lüdtke, O., & Robitzsch, A. (2019). Does personality become more stable with age? Disentangling state and trait effects for the big five across the life span using local structural equation modeling. *Journal of Personality and Social Psychology*, 116(4), 666–680. <https://doi.org/10.1037/pspp0000203>
- Wendt, L. P., Jankowsky, K., Zimmermann, J., Schroeders, U., Nolte, T., Fonagy, P., ... Olaru, G. (2021, January 8). Established self-report questionnaires of psychopathology can be efficiently summarized with HiTOP. <https://doi.org/10.31234/osf.io/j4gp8>
- Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, 44(2), 180–198. <https://doi.org/10.1016/j.jrp.2010.01.002>
- Zettler, I., Thielmann, I., Hilbig, B. E., & Moshagen, M. (2020). The nomological net of the HEXACO model of personality: A large-scale meta-analytic investigation. *Perspectives on Psychological Science*, 15(3), 723–760. <https://doi.org/10.1177/1745691619895036>