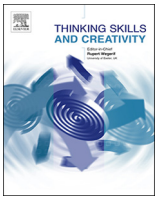
Thinking Skills and Creativity 36 (2020) 100659

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc



Psychometric properties of the Actively Open-minded Thinking scale

Eva M. Janssen[a,*], Peter P.J.L. Verkoeijen[b,c], Anita E.G. Heijltjes[c], Tim Mainhard[a], Lara M. van Peppen[b], Tamara van Gog[a]

[a] Department of Education, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, the Netherlands
[b] Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands
[c] Learning and Innovation Center, Avans University of Applied Sciences, Hogeschoollaan 1, 4818 CR Breda, the Netherlands

ARTICLE INFO

Keywords:
Dutch-language version of the Actively open-Minded Thinking scale
Thinking dispositions
Rational thinking
Reasoning and decision-making
Assessment
Mokken scale analysis

ABSTRACT

The Actively Open-minded Thinking scale (AOT; Stanovich & West, 2007) is a questionnaire that is used to measure the disposition towards rational thinking as a single psychological trait. Yet, despite its frequent use, also in abbreviated form, it is still unclear whether sumscores of the AOT can actually be used to order individuals on their disposition towards actively open-minded thinking and whether the questionnaire can be validly shortened. The present study aimed to obtain a valid and shorter AOT. We conducted Mokken scale analyses on the (Dutch) AOT using two samples of higher education students (N = 930; N = 509). Our analyses showed that none of the 41 items could discriminate sufficiently between respondents with varying latent trait levels. Furthermore, no item-set of the AOT could be obtained to validly order individuals on the assumed latent trait, which is a crucial assumption when using it in research. Consequently, it is questionable whether scores on the AOT provide insights into the concept it aims to measure.

1. Introduction

Baron (1991) introduced the concept actively open-minded thinking as an ideal standard of thinking, aimed at avoiding the tendency to reason based on intuitive heuristics and to focus instead on reflection about rules of inference. After Baron introduced the concept, the Actively Open-minded Thinking scale (AOT) developed by Stanovich and West (1997, 2007) has been widely used a measure of people's disposition towards rational thinking (see Table 1). The AOT has been shown to predict performance on, for example, critical thinking tests and is an important measure in reasoning and decision-making research (e.g., Heijltjes, Van Gog, Leppink, & Paas, 2014; Toplak, West, & Stanovich, 2011; West, Toplak, & Stanovich, 2008). The most widely used version of the AOT (Stanovich & West, 2007) consists of 41 statements in the form of Likert type items. Because this version takes substantial time to administer, it would be practical to obtain a valid shorter version (cf. the abbreviated versions to assess Need for Cognition, another widely used disposition scale; Cacioppo, Petty, & Feng Kao, 1984; Chiesi, Morsanyi, Donati, & Primi, 2018). In the literature, many different versions of the AOT have been applied, with item selections ranging from as few as 7 items (e.g., Haran, Ritov, & Mellers, 2013) to 47 items (e.g., Stanovich & West, 1997; see also Table 1).

* Corresponding author.
E-mail addresses: e.m.janssen@uu.nl (E.M. Janssen), p.p.j.l.verkoeijen@essb.eur.nl, ppjl.verkoeijen@avans.nl (P.P.J.L. Verkoeijen), aeg.heijltjes@avans.nl (A.E.G. Heijltjes), m.t.mainhard@uu.nl (T. Mainhard), vanpeppen@essb.eur.nl (L.M. van Peppen), t.vangog@uu.nl (T. van Gog).

https://doi.org/10.1016/j.tsc.2020.100659
Received 22 November 2019; Received in revised form 23 March 2020; Accepted 6 April 2020
Available online 18 April 2020

Table 1
Different versions of the Actively Open-minded Thinking scale (AOT) developed by Stanovich and West (2007) Applied in Studies from 2007 to 2019.

Studies	No. items	Response format	Language	Intermixed	AOT measure	Reported psychometrics
Deniz et al. (2008)	41	4-point scale	Turkish	No	Sumscore	α = .72
Stanovich and West (2008)	41	6-point scale	English	Yes	Sumscore	α = .84
West et al. (2008)	41	6-point scale	English	Yes	Standardized sumscore	SH = .76, α = .84
Elik, Wiener, and Corkum (2010)	41	6-point scale	English (?)	No	Sumscore (?)	α = .87
Deniz (2011)	41	5-point scale	English	No	Sumscore (?)	α = .80
Gerber and Scott (2011)	41	6-point scale	English	Yes	Standardized sumscore	SH = .75, α = .83
Toplak et al. (2011)	41	6-point scale	English	Yes	Sumscore	SH = .78, α = .81
Athanasiou and Papadopoulou (2012)	41	5-point scale	Greek	No (?)	Sumscore	α = .78
Athanasiou, Katakos, and Papadopoulou (2012)	n.r.	n.r.	Greek (?)	No (?)	Sumscore	α = .71
Beatty and Thompson (2012)	41	6-point scale	English (?)	No	Sumscore	n.r.
West, Meserve, and Stanovich (2012)	41	6-point scale	English	Yes	Sum score	α = .83
Haran et al. (2013)	7	7-point scale	English	No	n.r.	n.r.
Thompson et al. (2013)	41	6-point scale	English (?)	No	Sumscore	n.r.
Campitelli and Gerrans (2014)	15	6-point scale	English	No	Sumscore	n.r.
Heijltjes, Van Gog, and Paas (2014)	41	6-point scale	Dutch	Yes	Sumscore	α = .84
Heijltjes, Van Gog, Leppink et al. (2014)	41	6-point scale	Dutch	Yes	Sumscore	α = .82
Klaczynski (2014)	10	6-point scale	English	Yes	Composite score of AOT with 4 other scales	α = .78 (for composite score)
Pennycook, Cheyne, Barr, Koehler, and Fugelsang (2014)	32	n.r.	English (?)	Yes	Sumscore	α = .85
Shu and Townsend (2014)	8 (?)	n.r.	English	n.r.	n.r.	n.r.
Spadaccini and Esteves (2014)	41	6-point scale	English	No	Sumscore	α = .81
Swami, Voracek, Stieger, Tran, and Furnham (2014)	41	6-point scale	English	No	Averaged item score	α = .76
Thompson and Johnson (2014)	41	n.r.	English	No	Sumscore	n.r.
Toplak et al. (2014a)	30 (?)	4-point scale	English (?)	Yes	Standardized composite score of AOT with 1 other scale	SP = .60, α = .59 (for AOT)
Toplak et al. (2014b)	41	6-point scale	English	Yes	Sumscore	SH = .85, α = .86
Athanasiou and Papadopoulou (2015)	n.r.	n.r.	Greek and Turkish	No	Sumscore	α = .71 and α = .72
Baron, Scott, Fincher, and Emlen Metz (2015)	8	5-point scale	English	No	n.r.	α = .67
Heijltjes, Van Gog, Leppink, and Paas (2015)	41	6-point scale	Dutch	Yes	Sumscore	α = .70
Mellers et al. (2015)	7	7-point scale	English	n.r.	Averaged item score	α = .65
Pennycook, Fugelsang, and Koehler (2015)	41	n.r.	English (?)	Yes	Composite score of AOT with 1 other scale	α = .91 (for composite score)
Price, Ottati, Wilson, and Kim (2015)	Unclear	7-point scale (?)	English	Yes (?)	n.r.	n.r.
Jurkovič (2016)	Unclear	6-point scale	Slovak	Yes	Composite score of AOT with with new developed items	n.r.
Krumrei-Mancuso and Rouse (2016)	41	6-point scale	English	No	Sumscore	α = .92
Lean Keng and AlQudah (2017)	41	6-point scale	Arabic (?)	No	Averaged item score	α = .75
Toplak, West, and Stanovich (2017)	12	6-point scale	English	Yes	Sumscore	α = .71
Bautista, Misco, and Quaye (2018)	46	6-point scale	English	No	Sumscore	α = .81
Thompson, Pennycook, Trippas, and Evans (2018)	41	n.r.	English (?)	No	n.r.	n.r.

Note. This table provides an overview of scientific articles in which use of the Actively Open-minded Thinking scale (AOT; Stanovich & West, 2007) was reported and shows what kind of version was used. This overview was generated as follows: we surveyed the list of publications referring to Stanovich and West (2007) in Thomson Web of Science on February 7, 2019 and we included the articles that applied the AOT in some way in this table (i.e., theoretical or review articles, conference papers, and empirical articles referring but not applying the AOT were ignored). Studies are listed in chronological, then alphabetical order.
No. items shows number of selected items in the study; Intermixed shows whether the AOT items were intermixed with items of other questionnaires during the assessment; AOT measure shows in what way the AOT was included in the analyses; Reported psychometrics shows what psychometric statistics were reported on the AOT measure.
N r. = not reported in the manuscript; (?) = this was not explicitly mentioned but became implicitly clear from the manuscript; Unclear = the authors reported explicitly that they used a different item selection than the original version but did not report how many items they adopted; α = Cronbach's alpha; SH = split-half reliability (Spearman-Brown corrected).

Thus far, based on considerably high Cronbach's alpha's, most versions of the AOT including the original version, have been considered to be reliable. Despite the high Cronbach alpha's and the associations of the AOT scores with variables like critical thinking test performance, it is as yet unclear to what degree the items together reflect the single psychological trait of actively open-minded thinking. In other words, it is unclear what the internal validity of the AOT is and because of this, it is hard to substantively interpret correlations of the AOT with other variables. Thus, to further advance and strengthen research in the domain of reasoning and decision-making, it is important to investigate whether and to what degree the items included in the AOT measure the psychological trait actively open-minded thinking.

The little research that has been conducted on the internal validity of the AOT has used Factor Analysis (FA; Stanovich & West, 1997; Svedholm-Häkkinen & Lindeman, 2018). The results of these FAs, however, shed doubt on whether the AOT sumscores measure a single psychological trait and, thus, on whether AOT sumscores can be used to order individuals on the assumed latent trait; a finding we will elaborate on shortly. Furthermore, an assumption of FA is that the item scores are continuous and although this is a common assumption in social sciences research, psychometricians argue that Likert scale data are not continuous and should be treated as categorical data (Flora, LaBrish, & Chalmers, 2012; Jamieson, 2004; Liddell & Kruschke, 2018).

An alternative test theory approach that is suitable for categorical data is Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002). Mokken scale analysis is a non-parametric item response theory approach which tests whether a set of items can be used to order individuals on an assumed latent trait. Further benefits of this analysis are that it does not require multivariate normality or linear correlation between items, which are two assumptions of FA that are easily violated when working with Likert scale data (Flora et al., 2012; Jamieson, 2004; Liddell & Kruschke, 2018). Therefore, the present study applied Mokken scale analyses on (Dutch) AOT data (translated version used in Heijltjes, Van Gog, Leppink et al., 2014), to see whether we could obtain a shorter and valid version. In addition, for comparability with prior work by Stanovich and West (2007) and by Svedholm-Häkkinen and Lindeman (2018) we also conducted two Confirmatory Factor Analyses (CFAs), applying their proposed models.

1.1. Actively Open-minded Thinking

According to Baron (2008), good *thinking* consists of (1) "search that is thorough in proportion to the importance of the question, (2) confidence that is appropriate to the amount and quality of thinking done, and (3) fairness to other possibilities than the one we initially favor" (p. 200, italics added); *open-minded* refers to "the consideration of new possibilities, new goals, and evidence against possibilities that already seem strong" (p. 200, italics added); and *active* refers to not waiting for these things to happen but seeking them out. In his work, Baron argued that our thinking often deviates from the ideal of actively open-minded thinking, which leads to biases in our reasoning and decision-making. Think for example of our tendency to ignore evidence that goes against the conclusion we favor (i.e., confirmation bias).

Actively open-minded thinking is viewed as a thinking disposition. Thinking dispositions (or cognitive styles) are viewed as relatively stable psychological mechanisms that tend to generate characteristic behavioral tendencies and tactics (Stanovich, West, & Toplak, 2016). Thinking dispositions reflect people's goal management, epistemic values, and epistemic self-regulation. In their book "The rationality quotient", Stanovich et al. (2016), claim that, next to intelligence, thinking dispositions underlie rationality. They argue that – independent of cognitive ability – those who habitually seek various points of view or think extensively about a problem tend to display more rational behavior than those without such thinking dispositions. Psychologists have studied many thinking dispositions in relation to rationality. For instance, an individual's tendency to engage in and enjoy thinking, measured with the Need for Cognition scale, (NFC; Cacioppo & Petty, 1982; Cacioppo et al., 1984) has shown to be positively associated with rational thinking skills[1] after controlling for variance due to cognitive ability (Toplak & Stanovich, 2002; West et al., 2008). Other examples of thinking disposition questionnaires are the Rational-Experiential Inventory (Epstein, Pacini, Denes-Raj, & Heier, 1996) and the Consideration of Future Consequences scale (Strathman, Gleicher, Boninger, & Edwards, 1994). The disposition to think actively open-minded, however, is theorized to be the most central to rational thinking (Baron, 2008; Stanovich et al., 2016).

1.2. The Actively Open-minded Thinking scale

Baron (2008) mostly measured actively open-minded thinking qualitatively by assessing people's beliefs of good thinking, for example, through asking them to evaluate exemplars of the thinking of others. Inspired by Baron's work, Stanovich and West (1997, 2007) composed a questionnaire to measure actively open-minded thinking, the Actively Open-minded Thinking scale (AOT). The AOT consists of statements about thinking based on which participants rate their (dis)agreement on a Likert response format with six categories: 6 (agree strongly), 5 (agree moderately), 4 (agree slightly), 3 (disagree slightly), 2 (disagree moderately), 1 (disagree strongly). An example of such a statement is: "A person should always consider new possibilities."

For the first version in 1997, Stanovich and West (1997) composed 56 items distributed across eight subscales (Flexible Thinking, Openness Values, Dogmatism, Categorical Thinking, Openness-ideas, Absolutism, Superstitious Thinking, and Counterfactual Thinking). They found only three out of the eight subscales were to be reliable.[2] Moreover, a principal components analysis (PCA)

[1] In these studies, rational thinking is operationalized in ability to avoid bias in reasoning and decision-making measured with performance on so-called heuristics-and-biases tasks (West et al., 2008). These tasks measure whether someone is prone to a specific bias during a specific type of reasoning.

[2] Split-half reliability and Cronbach's alpha were, respectively, .49 and .50 for Flexible thinking; .73 and .71 for Openness values; .54 and .60 for

revealed that the first six subscale sumscores formed one component explaining most of the variance (40.8 %). Consequently, they excluded the subscales Superstitious Thinking and Counterfactual Thinking and computed a single composite score using the remaining subscales (i.e., summing the scores on the Flexible Thinking, Openness-Ideas, and Openness-Values scales and subtracting the sum of the Absolutism, Dogmatism, and Categorical Thinking scales), hereby treating the AOT as a unidimensional trait without subfactors. This score intended to order respondents on a scale ranging from "openness to belief-change and cognitive flexibility" (high scores) to "cognitive rigidity and resistance to belief change" (low scores). In the 1997 study, the Spearman-Brown corrected split-half reliability of the scale was .90 and Cronbach's α of the scale as a whole was .88.

Ten years after the first introduction of the AOT (Stanovich & West, 1997), Stanovich and West (2007) introduced a 41-item AOT, which from then became the most widely used version of the AOT. The scale consisted again of six subscales: four scales remained the same as in the first version (i.e., Flexible Thinking, Openness Values, Dogmatism, Categorical Thinking), but two scales (Openness-Ideas and Absolutism) were replaced with the subscales Belief Identification and Counterfactual thinking. A sumscore of the 41 items (after reverse scoring of 30 items) intended to order respondents on their disposition towards actively open-minded thinking. Again, the reliability for the total scale was good (split half reliability .75 and Cronbach's alpha .83). Because the subscale reliabilities were not reported, we assume that the new item selection was again intended to assess actively open-minded thinking as a unidimensional trait. However, an analysis to test this assumption was not reported.

1.3. Research using the Actively Open-minded Thinking scale

Since its introduction, numerous researchers have used the AOT. For example, multiple studies showed that the sumscores of this scale positively correlated with measures of rational thinking, with significant correlation coefficients ranging from .10 to .85 (e.g., Sá, West, & Stanovich, 1999; Sá, Kelley, Ho, & Stanovich, 2005; Sá & Stanovich, 2001; Toplak et al., 2011; West et al., 2008; Heijltjes, Van Gog, Leppink et al., 2014; Lean Keng & AlQudah, 2017; Svedholm-Häkkinen & Lindeman, 2018). A search in Web of Science (February 2019) indicated that the scientific papers introducing the first (Stanovich & West, 1997) and second version of the AOT (Stanovich & West, 2007) have been cited in 205 and 87 journal articles, respectively. We reviewed the 87 studies citing the 2007 version (currently the most widely used version) and found that 36 had adopted (a part of) the AOT as a measure (see Table 1). Researchers used it – often in combination with other disposition questionnaires – as a predictor of reasoning and decision-making (e.g., performance on rational, scientific, or analytic reasoning tasks or political choices), epistemic beliefs (e.g., evolutionary theory acceptance, belief in conspiracy theories, or religiosity), or behavior (e.g., being a gambler or showing adaptive teaching behavior). Studies varied widely, however, in the way the scale was administered. Some relatively small differences concerned the response formats (4-point to 7-point Likert scales) and whether the items were intermixed with other (disposition) questionnaires. A more important difference concerned the item selection. Within those 36 studies, 12 studies used a different item selection than the original 41-item AOT (see Table 1). In addition, in their book 'The Rationality Quotient' Stanovich et al. (2016) introduced a 30-item version and a 16-item short-form. Most studies did report a sufficient reliability for the total scale, but as listed in Table 1, none tested the factor structure. Hence, it is still an open question whether and to what degree the AOT from 2007 and the newer versions measure actively open-minded thinking as a unidimensional trait.

Recently, Svedholm-Häkkinen and Lindeman (2018) noted that it is not clear whether the AOT is unidimensional or multidimensional because the PCA on the first 47-item AOT (Stanovich & West, 1997) was run on sumscores of the subscales rather than single items and because subsequent studies reported reliability measures for the scale as a whole only. Svedholm-Häkkinen and Lindeman (2018) aimed to develop a reliable, valid, shorter AOT and investigated whether the AOT was multidimensional or not. To this end, they conducted FAs in four separate samples ($N = 2735$, $N = 458$, $N = 102$, and $N = 50$) who had completed a Finnish version of the 41-item AOT (Stanovich & West, 2007). A 17-item version was sufficient to obtain a good reliability (Cronbach's alpha) for the total scale and to obtain correlations with variables assessing other thinking dispositions, social competence, and supernatural beliefs. However, their results also showed that the AOT was not unidimensional. They compared five different factor models and concluded that four intercorrelated subfactors (Dogmatism, Fact resistance, Liberalism, and Belief personification) described the data best. Neither a model with a higher-order factor (i.e., representing active open-mindedness) explaining the common variance in the four subscales, nor a single factor solution described the data adequately, which suggests that AOT sumscores cannot be used to validly order individuals on the assumed psychological trait of active open-mindedness. In addition, just as in the study by Stanovich and West (1997), the four subscales were only marginally or not reliable[3] .

1.4. The present study

In sum, despite its frequent use, previous studies have not yet demonstrated that the AOT is a valid measurement instrument to order individuals on actively open-minded thinking. High reliability values indicate that the *observed* AOT sumscores can classify individuals low to high. In addition, a positive correlation of observed sumscores with an external predictor shows that a high AOT

(*footnote continued*)

Dogmatism; not reported for Categorical thinking because the scale consisted of only 2 items; .73 and .77 for Openness-ideas; and .69 and .64 for Absolutism, not reported for Counterfactual thinking because the scale consisted of only 2 items; .73 and .73 for Superstitious thinking (see Stanovich & West, 1997).

[3] Cronbach's alpha in Study 1 was .67 for Dogmatic thinking; .67 for Fact resistance; .43 for Liberalism; .56 for Belief personification.

sumscore is likely to go together with a high score on a relevant other variable. When an AOT sumscore is computed and used in analyses, the implicit assumption is that this set of items can be used to order individuals on the assumed latent trait representing active open mindedness. Both reliability and correlational analyses, however, do not answer the question of whether the items used (i.e., the specific AOT sumscore), measure a single latent trait. Psychometric validation of the AOT scale, therefore, requires testing this assumption by assessing whether the responses of individuals to each item can be described as a function of a single latent trait (i.e., internal validity).

Svedholm-Häkkinen and Lindeman (2018) tested this assumption using FA; however, their FAs suggested that the items they included did not reflect a unidimensional trait (i.e., no higher-order factor). Hence, a sumscore of their 17-item solution cannot be used to validly order individuals on the latent trait of actively open-minded thinking either, as it is unclear what concept a sumscore on this abbreviated version reflects. To illustrate our point, imagine that we measured four traits with a questionnaire: social economic status (SES) with 4 items, work satisfaction with 4 items, motivation to eat healthy with 4 items, and engagement in politics with 5 items. If we validly measured the four traits and subjected all 17 items to a FA, one would expect to find four intercorrelated factors without a higher-order factor (i.e., similar to the factor structure as Svedholm-Häkkinen & Lindeman, 2018). However, even if the reliability was sufficient, a sumscore of those 17 items cannot be interpreted meaningfully, because a sum of a person's SES and motivation to eat healthy cannot easily be interpreted as a character trait of a person (i.e., the scales do not form a higher-order unidimensional trait). Also if this sumscore would correlate with other variables (e.g., mental health or having debts), this still does not allow for interpreting the sumscore as a single latent trait.

The aim of this study was to re-examine the validity of all 41 items of the AOT developed by Stanovich and West (2007), to see whether we could develop a valid shorter version that allows for ordering participants on the assumed latent trait. To this end, we used Mokken scale analysis. Mokken scale analysis is a non-parametric item response theory approach, which tests whether a set of items can be used to order individuals on an assumed latent trait. Moreover, Mokken scale analysis has advantages over the commonly conducted FAs, when analyzing data based on Likert type scales. An important advantage is that Mokken scale analysis is suitable for categorical data whereas FA requires data at the interval level. In the field of psychometrics it is argued that treating Likert scale data as interval data can be problematic (Liddell & Kruschke, 2018). When treating Likert scale data instead as being ordinal, it is technically not possible to test for the normality assumption of FA because the difference between two successive values cannot be quantified. Even if one would still assess normality with ordinal data, Likert-scale items typically indicate skewed or polarized distributions (Jamieson, 2004; Liddell & Kruschke, 2018). An additional advantage of Mokken scale analysis in this respect, is that it does not require multivariate normality or linear correlations between items. In the present study, we conducted an exploratory Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002) on the Dutch version of the AOT (Heijltjes, Van Gog, Leppink et al., 2014; Stanovich & West, 2007) using two samples of higher education students (N = 930 and N = 509) to see whether we could obtain a valid shorter version of the AOT. In addition, for comparability with prior work by Stanovich and West (2007) and by Svedholm-Häkkinen and Lindeman (2018) we conducted two CFAs representing their proposed models.

2. Method

All materials, datasets, R-code, and output are stored on an Open Science Framework (OSF) page for this project, see https://osf.io/4hxzu/.

2.1. Participants and procedure

We obtained anonymized AOT datasets from a Dutch University of Applied Sciences[4], where the AOT (see next section) was filled out on a computer by first year students, as part of a critical thinking course they were enrolled in. It was not possible to skip questions and only fully completed questionnaires could be submitted, so there was no missing data.

We repeated the same Mokken scale analyses on two datasets to see whether we obtained similar results. Dataset A (N = 930) was a merged dataset that consisted of 460 students in the economics and business domain (data collected in 2014) and 470 students in the health care domain (data collected in 2016). Age and sex were indicated by 908 participants, whose mean age was 18.84 years (SD = 2.30) and 55 % of whom were female. Dataset B (N = 509) was a merged dataset that consisted of 257 students in the marketing and business management domain (data collected in 2017) and 252 students in the health care domain (data collected in 2017). Age and sex were indicated by 506 participants, whose mean age was 18.82 years (SD = 2.46) and of whom 50 % were female.

2.2. Actively Open-minded Thinking scale

We used a Dutch translation of the original 41-item version of the AOT (Stanovich & West, 2007). In a previous study, Heijltjes, Van Gog, Leppink et al. (2014) made the Dutch translation, which was checked by two persons one of whom was a native English speaker. In line with the original scale, the response format consisted of six answering categories: Strongly agree (6), Moderately

[4] The Dutch education system distinguishes between higher professional education offered by universities of applied sciences (Bachelor, Master), and academic education offered by academic universities (Bachelor, Master, PhD, with the PhD being an additional four-year trajectory after a Master degree).

agree (5), Slightly agree (4), Slightly disagree (3), Moderately disagree (2), and Strongly disagree (1). We reverse scored 30 items so that for all items a higher score indicated a stronger disposition towards actively open-minded thinking[5] (for all items, see https://osf.io/4hxzu/).

2.3. Analyses

2.3.1. Mokken scale analyses

We conducted an exploratory Mokken scale analysis (Mokken, 1971; Molenaar & Sijtsma, 2000) aiming to extract a valid shorter version of the AOT from the total item pool. The Mokken scale analysis was performed using the Monotone Homogeneity Model and the Automated Item Selection Procedure algorithm from the 'Mokken' package in R (R Development Core Team, 2008; Van der Ark, 2007). The Automated Item Selection Procedure in Mokken scale analysis partitions a set of items from an item pool into one or more scales. Items included in such a scale need to have sufficient discriminative power. Items that do not, or only very weakly, discriminate between persons with varying latent-trait levels are left unscalable (Sijtsma & Molenaar, 2002).

In contrast to factor analysis, Mokken scale analysis requires only few assumptions and is, therefore, robust to problems concerning the distribution of the underlying data. First, the model underlying Mokken scale analysis assumes unidimensionality, which means that all items in a particular test measure the same latent trait (Sijtsma & Molenaar, 2002). Second, the underlying model assumes local independence, which means that a person's response to an item is not influenced by his or her responses to the other items in the test, given the underlying latent trait. However, if students gain knowledge during the test, which they can use to answer the next items in the very same test, the assumption of local independence is violated. The third and final assumption is that the probability of answering the item correctly (or, in case of polytomous items, the probability of agreeing to the item) increases or stays the same as the ability level increases or, put more technically, that the response functions of the items (IRFs) are monotonically nondecreasing (Sijtsma & Molenaar, 2002). Furthermore, Mokken scale analysis is suitable for the analysis of categorical data. The AOT items all have six ordered response categories, ranging from "disagree strongly" to "agree strongly". Inspecting the frequency distributions of the 41 AOT items indicated that for some items the distribution was skewed (for these results, see https://osf.io/4hxzu/). As such, Mokken scale analysis was most suitable for our AOT data as it does not assume the data to be normally distributed.

Three scalability coefficients were used to determine whether or not items formed a scale, and as diagnostics to assess the strength of the scales (Kuijpers, 2015): (1) item-pair scalability coefficient H_{ij}, which expresses the strength of the association between items i and j given their marginal distributions; (2) item scalability coefficient H_j, which expresses how well item j fits with the other items in a test, and also indicates the extent to which item j discriminates between respondents (Sijtsma & Molenaar, 2002, p. 66); and (3) total-scale scalability coefficient H, which expresses the degree to which respondents can be ordered by means of a set of items (Sijtsma & Molenaar, 2002, pp. 36, 39). A set of items can be used to order individuals on the assumed latent trait if (1) all $H_{ij} \geq 0$ (i.e., the underlying model assumes positive inter-item covariances) and (2) if $H_j > c > 0$ for all j. The latter indicates that all item scalability coefficients should be at least positive, and rather above a positive value c (by default set to .3), such that non-discriminating items or only weakly discriminating items are excluded from the scale. As follows from these two criteria, the value of the total-scale scalability coefficient H should be at least .3 (Kuijpers, Van der Ark, & Croon, 2013; Mokken, 1971; Molenaar & Sijtsma, 2000). H-values lower than .3 are regarded as indicating that the set of items is poorly scalable. Finally, note that sufficient scalability coefficients imply that a set of items can be used to order individuals on an assumed latent trait. Obtaining sufficient scalability coefficients does, however, not automatically imply that this set of items is measuring a unidimensional construct (Smits, Timmerman, & Meijer, 2012). To gain insight into the dimensionality of a scale, factor modeling is a more suitable method.

2.3.2. Confirmatory factor analyses

To gain insight into the dimensionality of the AOT and for comparability with prior work by Stanovich and West (2007) and by Svedholm-Häkkinen and Lindeman (2018), we also ran two CFAs on dataset A and B. First, we ran a CFA on the model proposed by Stanovich and West (2007): a one factor with the 41 items as indicators of a single trait. Second, we ran a CFA based on the final model proposed by Svedholm-Häkkinen and Lindeman (2018): a 17-item version with four intercorrelated factors without one higher order factor. We used the '*Lavaan*' package in R (R Development Core Team, 2008; Rosseel, 2012) with robust weighted least squares (WLSMV) as estimation method. This estimator is seen as most suitable for categorical data (Brown, 2006). To be fully consistent with Svedholm-Häkkinen and Lindeman (2018), we also ran the CFAs with ML estimation, yielding highly similar results (for these results, see https://osf.io/4hxzu/).

We followed the guidelines by Hu and Bentler (1999) to examine the model fit. Hu and Bentler (1999) argue that values close to .95 for the Comparative Fit Index (CFI) and the Tucker Lewis Index (TLI), in combination with values close to .06 and .08 for the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR), respectively, are needed to conclude that there is a relatively good fit between the hypothesized model and the observed data. We used the standardized factor loadings to determine whether test items could discriminate between respondents with varying trait levels (Brown, 2006). Values of the standardized factor loadings should be at least .4 to be sufficient.

[5] The students additionally completed a Dutch translation of 18-item (short form) Need for Cognition questionnaire (NFC; Cacioppo et al., 1984) that we intended to use as criterion variable to assess the validity of our newly obtained AOT in subsequent analyses. However, since our results did not yield a useful item selection for the AOT (see Results section), we did not use this variable in further analyses.

Table 2
Clusters, Item Scalability Coefficients (Hj), and the Corresponding Standard Errors for the 18 Scalable Items Selected by the Automated Selection Procedure on Dataset A (N = 930).

Item		Cluster	M^a	*SD*	H_j	*SE*
5.	There are two kinds of people in this world: those who are for the truth and those who are against the truth. (R)	C1	4.10	1.30	0.376*	0.029
8.	I think there are many wrong ways, but only one right way, to almost anything. (R)	C1	5.00	0.98	0.325	0.032
17.	There are basically two, kinds of people in this world, good and bad. (R)	C1	4.63	1.34	0.369*	0.028
4.	A person should always consider new possibilities.	C2	5.01	0.78	0.310	0.034
37.	Beliefs should always be revised in response to new information or evidence	C2	4.50	0.85	0.351	0.032
41.	People should always take into consideration evidence that goes against their beliefs.	C2	4.65	0.89	0.350	0.034
15.	It is important to persevere in your beliefs even when evidence is brought to bear against them. (R)	C3	3.90	1.22	0.369	0.036
19.	Certain beliefs are just too important to abandon no matter how good case can be made against them. (R)	C3	3.11	1.13	0.369	0.036
25.	My beliefs would not have been very different if I had been raised by a different set of parents. (R)	C4	3.89	1.33	0.356	0.037
28.	Even if my environment (family, neighborhood, schools) had been different, I probably would have the same religious views. (R)	C4	2.85	1.38	0.356	0.037
23.	I believe that loyalty to one's ideals and principles is more important than "open-mindedness". (R)	C5	4.46	0.97	0.347	0.039
33.	One should disregard evidence that conflicts with your established beliefs. (R)	C5	4.82	0.89	0.347	0.039
11.	There are a number of people I have come to hate because of the things they stand for. (R)	C6	4.10	1.34	0.345	0.031
31.	My blood boils over whenever a person stubbornly refuses to admit he's wrong. (R)	C6	2.92	1.30	0.345	0.031
13.	No one can talk me out of something I know is right. (R)	C7	3.00	1.21	0.335	0.036
14.	Basically, I know everything I need to know about the important things in life. (R)	C7	4.32	1.17	0.335	0.036
16.	Considering too many different opinions often leads to bad decisions. (R)	C8	3.93	1.16	0.331	0.036
35.	A group which tolerates too much difference of opinion among its members cannot exist for long. (R)	C8	4.06	1.21	0.331	0.036

Note. R = reverse scored item.
[a] Higher = more actively open-minded thinking.
* H_j significantly above. 3 with $p < .05$.

3. Results

3.1. Mokken scale analyses

The Mokken scale analysis performed on the first dataset (A) showed that no subset of items could be constructed that validly order individuals on the assumed latent trait. None of the 41 items could discriminate sufficiently between respondents with varying latent trait levels, all $H_j s \leq 0.182$, $H = .105$ (for all 41 $H_j s$ and the item-pair scalability coefficients, see https://osf.io/4hxzu/). Furthermore, the explorative analyses indicated that 18 out 41 items formed eight separate scales with each two to three items at most. The remaining 23 items were left unscalable, that is, the items were not or even more poorly discriminating and/or covaried negatively with items included in one of the eight scales. Table 2 shows the eight scales and the item scalability coefficients (H_j) with the corresponding standard errors for the 18 scalable items (for the item-pair scalability coefficients, see https://osf.io/4hxzu/). Only two items had a H_j that was significantly above .3 and none of the scales consisted entirely of items with H_j significantly > .3. With regard to the scales' total scalability coefficients, only the first cluster (C1) had a coefficient significantly > .3, C1: $H = .359$, $SE = .026$; C2: $H = .338$, $SE = .030$; C3: $H = .369$, $SE = .036$; C4: $H = .356$, $SE = .037$; C5: $H = .347$, $SE = .039$; C6: $H = .345$, $SE = .031$; C7: $H = .335$, $SE = .036$; C8: $H = .331$, $SE = .036$.

In addition to the finding that the scales did not discriminate between respondents, the scales appeared to be unreliable, C1: latent class reliability coefficient (LCRC; Van der Ark, Van der Palm, & Sijtsma, 2011; Van der Palm, Van der Ark, & Sijtsma, 2014) = .55, Cronbach's alpha (α) = .58; C2: LCRC = .51, $\alpha = .55$; C3: LCRC = .25, $\alpha = .50$; C4: LCRC = .24, $\alpha = .49$; C5: LCRC = .23, $\alpha = .46$; C6: LCRC = .24, $\alpha = .49$; C7: LCRC = .23, $\alpha = .47$; C8: LCRC = .24, $\alpha = .48$.

Findings for dataset B were more or less similar. Again, none of the 41 items could discriminate sufficiently between respondents with varying latent trait levels, all $H_j s \leq .195$, $H = .100$ (for all 41 $H_j s$ and the item-pair scalability coefficients, see https://osf.io/4hxzu/). The explorative analyses identified seven separate scales with five items in the first scale and two items in each of the other six scales. The 24 remaining items were left unscalable. Table 3 shows the seven scales and the item scalability coefficients (H_j) with the corresponding standard errors for the 17 scalable items (for the item-pair scalability coefficients, see https://osf.io/4hxzu/). Again, only two items had a H_j that was significantly above .3 and none of the scales consisted entirely of items with H_j significantly > .3. With regard to the scales' total scalability coefficients, none of the clusters had a coefficient significantly > .3, C1: $H = .335$, $SE = .027$, C2: $H = .357$, $SE = .051$., C3: $H = .350$, $SE = .049$, C4: $H = .338$, $SE = .046$, C5: $H = .312$, $SE = .049$, C6: $H = .312$, $SE = .045$, C7: $H = .300$, $SE = .048$.

In addition to the finding that the scales did not discriminate between respondents, the scales appeared to be unreliable, C1: LCRC = .67, $\alpha = .66$; C2: LCRC = .23, $\alpha = .47$; C3: LCRC = .24, $\alpha = .48$; C4: LCRC = .24, $\alpha = .48$; C5: LCRC = .22, $\alpha = .44$; C6: LCRC = .22, $\alpha = .43$; C7: LCRC = .22, $\alpha = .44$.

Thus, the results of the Mokken scale analyses on both datasets A and B suggested that the 41 items together could not discriminate sufficiently between respondents with varying latent trait levels. Furthermore, no item-set of the AOT could be obtained to validly order individuals on the assumed latent trait. The item scales found did not have sufficient discriminative power and had a poor reliability. Moreover, it has been argued that using many subscales with only two or three items can have a negative impact on

Table 3
Clusters, Item Scalability Coefficients (Hj), and the Corresponding Standard Errors for the 17 Scalable Items Selected by the Automated Selection Procedure on Dataset B (N = 509).

Item		Cluster	M^a	*SD*	H_j	*SE*
5.	There are two kinds of people in this world: those who are for the truth and those who are against the truth. (R)	C1	3.93	1.30	0.315	0.034
8.	I think there are many wrong ways, but only one right way, to almost anything. (R)	C1	5.02	0.86	0.373*	0.031
17.	There are basically two, kinds of people in this world, good and bad. (R)	C1	4.49	1.38	0.361*	0.030
24.	Of all the different philosophies which exist in the world there is probably only one which is correct. (R)	C1	5.12	0.89	0.316	0.034
33.	One should disregard evidence that conflicts with your established beliefs. (R)	C1	4.73	0.92	0.309	0.034
6.	Changing your mind is a sign of weakness. (R)	C2	4.96	1.03	0.357	0.051
38.	I think that if people don't know what they believe in by the time they're 25, there's something wrong with them. (R)	C2	4.42	1.21	0.357	0.051
25.	My beliefs would not have been very different if I had been raised by a different set of parents. (R)	C3	3.88	1.29	0.350	0.049
28.	Even if my environment (family, neighborhood, schools) had been different, I probably would have the same religious views. (R)	C3	2.99	1.43	0.350	0.049
37.	Beliefs should always be revised in response to new information or evidence.	C4	4.47	0.87	0.338	0.046
41.	People should always take into consideration evidence that goes against their beliefs.	C4	4.58	0.91	0.338	0.046
21.	It is a noble thing when someone holds the same beliefs as their parents. (R)	C5	4.00	1.05	0.312	0.049
22.	Coming to decisions quickly is a sign of wisdom. (R)	C5	4.33	1.12	0.312	0.049
13.	No one can talk me out of something I know is right. (R)	C6	2.99	1.21	0.312	0.049
14.	Basically, I know everything I need to know about the important things in life. (R)	C6	4.33	1.14	0.312	0.049
35.	A group which tolerates too much difference of opinion among its members cannot exist for long. (R)	C7	4.08	1.18	0.300	0.048
39.	I believe letting students hear controversial speakers can only confuse and mislead them. (R)	C7	4.22	1.01	0.300	0.048

Note. R = reverse scored item.
[a] Higher = more actively open-minded thinking.
* H_j significantly above. 3 with $p < .05$.

the reliability, validity and measurement precision of a scale (Kruyen, Emons, & Sijtsma, 2013; Kruyen, Emons, & Sijtsma, 2012; Mellenbergh, 1996; Reise & Waller, 2009).

3.2. Confirmatory factor analyses

We first conducted a CFA on dataset A and B, testing the one-factor model on all 41 items as proposed by Stanovich and West (2007). For dataset A, we obtained mixed results on the model fit indices. For the absolute fit indices, Chi-square indicated a poor fit, χ^2 (779) = 3508.85, $p < .001$, which could be expected given the large sample size. RMSEA and SRMR, on the other hand, showed an acceptable fit, indicating that there was an acceptable discrepancy between hypothesized model (with optimal parameter estimates) and the actually obtained sample data (covariance matrix), RMSEA = 0.061; SRMR = 0.070. The incremental fit indices (analogous to R^2), however, showed poor fit, indicating that the improved data fit by the tested one-factor model was only marginal when compared to the data fit of the null model (in which all the observed variable are uncorrelated), CFI = 0.686; TLI = 0.669. Following the guidelines of Hu and Bentler (1999), we concluded that the model did not describe the data adequately. Furthermore, 25 out of 41 items had small standardized factor loadings (< .4; for all factor loadings, see https://osf.io/4hxzu/), indicating that those items could not discriminate between respondents with varying trait levels (Brown, 2006). In line with Stanovich and West (2007), the scale as a whole was reliable, $\alpha = .81$.

This same model tested on dataset B was over-identified, which indicates that the model should not be interpreted. Moreover, WLSMV estimator could therefore not be used to compute robust standard errors and the adjusted test statistics. The model's parameters could only be estimated using diagonally weighted least squares (DWLS). These results showed more or less similar estimates as found for Dataset A using the WLSMV estimator, χ^2 (779) = 2628.24, $p < .001$; RMSEA = .068; SRMR = .074; CFI = 0.829; TLI = 0.820. Also, again 25 items had small standardized factor loadings (< .4) and the Cronbach's alpha for the total scale was .80.

Next, we conducted a CFA on dataset A and B, testing the intercorrelated four-factor model without a higher-order factor on the 17-item AOT as proposed by Svedholm-Häkkinen and Lindeman (2018). Both for dataset A and B, we obtained mixed results. For dataset A, the absolute fit indices indicated an acceptable fit, $\chi^2(113) = 763.84$, $p < .001$; RMSEA = 0.079; SRMR = 0.066, whereas the incremental fit indices indicated a poor data fit, CFI = 0.782; TLI = 0.737. Hence, this model also did not describe the data adequately. Additionally, five out of 17 items had small standardized factor loadings (< .4; for all factor loadings, see https://osf.io/4hxzu/). The scale as a whole had a Cronbach's alpha of .67, and the subscales Dogmatism, Fact Resistance, Liberalism, and Belief Personification had alphas of .53, .56, .32 and .51 respectively. For dataset B, we obtained more or less similar results, $\chi^2(113) = 546.30$, $p < .001$; RMSEA = .087, $p < .001$; SRMR = .079; CFI = 0.763; TLI = 0.714. Six items had small standardized factor loadings (< .4) and the Cronbach's alpha for the total scale was .66. The subscale alphas were .60, .51, .24, .46, for Dogmatism, Fact Resistance, Liberalism, and Belief personification, respectively.

In sum, both the Mokken scale analyses and the one-factor CFA did not yield an item set that could be used to validly order individuals on the latent trait actively open minded thinking.

4. Discussion

The aim of this study was to obtain a valid shorter version of the AOT developed by Stanovich and West (2007) that could be used to order individuals on the latent trait actively open-minded thinking. Our results did not provide support for the hypothesis that either the 41-item AOT or a subset of items would measure actively open-minded thinking as a single latent trait. The Mokken scale analyses performed on two large datasets of Dutch first-year higher professional education students showed that none of the items discriminated very well between students on the (assumed) latent trait. In addition, no adequate AOT subscales could be identified. These findings imply that – for the studied population – sumscores on the AOT do not provide insight into the concept it aims to measure.

4.1. Relating the current results to previous findings

Sumscores on the AOT are widely used in, for example, correlational analyses. When one computes a sumscore and assumes that it provides insights into the construct "actively open-minded thinking", one assumes that all items load on the same assumed latent trait. However, the evidence so far, including our results, do not support this assumption. Together, the results of our Mokken scale analyses, our one-factor CFA, and the results of Svedholm-Häkkinen and Lindeman (2018) indicated that the AOT is not measuring one unitary trait. This renders the reported reliabilities for the scale as a whole (see Table 1) rather meaningless, as Cronbach's alpha assumes a unidimensional construct.

In addition to the fact that the scale does not measure a unidimensional trait, we also found no evidence for meaningful subscales. Here, our results differ somewhat from Svedholm-Häkkinen and Lindeman (2018), who found that a 17-item version of the AOT measured four separate subscales. Our Mokken scale analyses indicated that more than half of the 41 items were left unscalable (i.e., could not be included in a subscale) and that none of the (very small) subscales that were formed had sufficient discriminative power. Hence, the items included in a subscale could not order participants with varying levels of the latent trait that the scale was potentially measuring. Furthermore, the formed subscales were not reliable. Our CFA testing the four-factor model (without one higher-order factor) proposed by Svedholm-Häkkinen and Lindeman (2018) did not describe the data adequately. We obtained acceptable values for the absolute fit indices, but a poor values for the incremental fit indices. Hence, the tested four-factor model fitted acceptably with the obtained sample data but the model fitted the data only slightly better than the worst possible model would do. Furthermore, five (dataset A) and six (dataset B) of the 17 items had low factor loadings and could thus not discriminate between participants. Also note that both in our study and in the study by Svedholm-Häkkinen and Lindeman (2018) low reliabilities for the four subscales were obtained. Hence, there is still no convincing evidence that scores on the subscales can be interpreted meaningfully. One possible explanation for the divergent results with Svedholm-Häkkinen and Lindeman (2018) regarding the CFAs may be that, the Likert type AOT items are not suitable for FA and therefore do not yield robust results across studies (Magidson & Vermunt, 2003). A more likely explanation, however seems that the AOT items do not sufficiently measure what they intend to. Taking the Mokken scale and FAs together, it seems that the AOT items included in the available studies so far are not measuring a single psychological trait actively open-minded thinking nor any subtraits. The construct validity and content validity of the items should be improved in order to obtain a valid measurement instrument.

If it is unclear what the sumscore on the AOT represents, it is also unclear how the correlations that previous studies found between the AOT and other variables (such as other dispositions like the tendency to enjoy and engage in thinking, measured with the NFC, or performance on critical thinking tests Heijltjes, Van Gog, Leppink et al., 2014; Svedholm-Häkkinen & Lindeman, 2018; Toplak, West, & Stanovich, 2014; Toplak, West, & Stanovich, 2014) should be interpreted. The correlations between the AOT and other thinking dispositions may mean that some AOT items measure more or less the same as some items from other disposition questionnaires, and that the AOT sumscores therefore correlate with these variables (e.g., an item in the AOT is 'If I think longer about a problem I will be more likely to solve it' and an item in the Need for Cognition scale is 'I would prefer complex to simple problems'). It may also be that both the AOT and its criterion variables (e.g., rational reasoning) implicitly measure something else that we are not aware of and that this causes a correlation (cf. the third variable problem).

4.2. Limitations and further research

To our knowledge, this is the first study that investigated the psychometric properties of the AOT using Mokken scale analysis, which can be considered to be more suitable than the more commonly used FAs because it is suitable for the categorical responses to the AOT items and robust to violations of multivariate normality and linear correlation between items (Flora et al., 2012; Jamieson, 2004; Liddell & Kruschke, 2018; Mokken, 1971; Sijtsma & Molenaar, 2002). Nevertheless, two potential limitations of our study should be noted. First, it is possible that all participants in our study sample were very strong actively open-minded thinkers (i.e., relatively high average item scores items and therefore quite homogeneous), resulting in little or no variance in item scores. However, based on the items' distributions and the range of item scores, we consider both study samples sufficiently heterogenous for testing the items' scalability (for these results, see https://osf.io/4hxzu/). In addition, participants in our sample had a rather similar total score on the 41-item AOT (M = 171.8, SD = 15.2) compared to the Stanovich and West (2007) that introduced this version of the AOT (M = 170.7, SD = 18.2). A second limitation is that our analyses were conducted on the Dutch translation of the AOT. To our knowledge, none of the translated versions, including ours, have been compared to data on the English version. Hence, it remains an open question to what extent findings obtained with the translated scales apply to the original English AOT. However, on theoretical grounds we see no reason to expect any strong translation effects. Moreover, the results of previous studies using translated versions

of the AOT seem compatible with the results of studies using the English version. That is, they showed comparable descriptive statistics (after correcting for number of included items and/or response format) and similar correlations of AOT scores with other variables (e.g., Deniz, Donnelly, & Yilmaz, 2008; Heijltjes, Van Gog, Leppink et al., 2014; Stanovich & West, 2007; Svedholm-Häkkinen & Lindeman, 2018). It should also be noted that investigating whether translated AOT scales are invariant to the English scale will be quite challenging as long as the factor structure is unclear. Nevertheless, based on these considerations, we cannot fully rule out the possibility that the current results were somehow affected by the fact that we used a Dutch translation instead of the English version. Therefore, it would be interesting to replicate our Mokken scale analyses with other datasets on the English version of the AOT.

4.3. Conclusion

To conclude, the results of our study suggest that there is no item set of the 41 item version of the AOT that can be used to validly order individuals on their ability to think active open-mindedly, which is a crucial assumption when using it in research. Consequently, it is questionable whether scores on the AOT provide insights into the concept it aims to measure. If the results of the present Mokken scale analyses would replicate with English AOT data, this would be a strong argument for starting the process of (re-)designing a new scale to measure actively open-minded thinking or to consider alternative measures of thinking dispositions.

Open science framework

All materials, datasets, R-code, and output are stored on an Open Science Framework (OSF) page for this project, see https://osf.io/4hxzu/.

CRediT authorship contribution statement

Eva M. Janssen: Conceptualization, Methodology, Formal analysis, Writing - original draft. **Peter P.J.L. Verkoeijen:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Anita E.G. Heijltjes:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision. **Tim Mainhard:** Writing - review & editing, Supervision. **Lara M. van Peppen:** Writing - review & editing. **Tamara van Gog:** Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

None.

Acknowledgements

This work was supported by the Netherlands Organization for Scientific Research under project number 409-15-203. The authors would like to thank Renske Kuijpers for her assistance with the data analysis.

References

Athanasiou, K., & Papadopoulou, P. (2012). Conceptual ecology of the evolution acceptance among Greek education students: Knowledge, religious practices and social influences. *International Journal of Science Education, 34*, 903–924. https://doi.org/10.1080/09500693.2011.586072.

Athanasiou, K., & Papadopoulou, P. (2015). Evolution theory teaching and learning: What conclusions can we get from comparisons of teachers' and students' conceptual ecologies in Greece and Turkey? *EURASIA Journal of Mathematics, Science and Technology Education, 11*, 841–853. https://doi.org/10.12973/eurasia.2015.1443a.

Athanasiou, K., Katakos, E., & Papadopoulou, P. (2012). Conceptual ecology of evolution acceptance among Greek education students: The contribution of knowledge increase. *Journal of Biological Education, 46*, 234–241. https://doi.org/10.1080/00219266.2012.716780.

Baron, J. (1991). Beliefs about thinking. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.). *Informal reasoning and education* (pp. 169–186). Hillsdale. https://www.sas.upenn.edu/-baron/papers/voss.pdf.

Baron, J. (2008). *Thinking and deciding* (4th ed.). Cambridge University Press.

Baron, J., Scott, S., Fincher, K., & Emlen Metz, S. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*, 265–284. https://doi.org/10.1016/j.jarmac.2014.09.003.

Bautista, N., Misco, T., & Quaye, S. J. (2018). Early childhood open-mindedness: An investigation into preservice teachers' capacity to address controversial issues. *Journal of Teacher Education, 69*, 154–168. https://doi.org/10.1177/0022487117702575.

Beatty, E. L., & Thompson, V. A. (2012). Effects of perspective and belief on analytic reasoning in a scientific reasoning task. *Thinking & Reasoning, 18*, 441–460. https://doi.org/10.1080/13546783.2012.687892.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* Guilford Press.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116–131. https://doi.org/10.1037/0022-3514.42.1.116.

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of Need for Cognition. *Journal of Personality Assessment, 48*, 306–307. https://doi.org/10.1207/s15327752jpa4803_13.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition, 42*, 434–447. https://doi.org/10.3758/s13421-013-0367-9.

Chiesi, F., Morsanyi, K., Donati, M. A., & Primi, C. (2018). Applying item response theory to develop a shortened version of the Need for Cognition scale. *Advances in Cognitive Psychology, 14*, 75–86. https://doi.org/10.5709/acp-0240-z.

Deniz, H. (2011). Examination of changes in prospective elementary teachers' epistemological beliefs in science and exploration of factors meditating that change. *Journal of Science Education and Technology, 20*, 750–760. https://doi.org/10.1007/s10956-010-9268-x.

Deniz, H., Donnelly, L. A., & Yilmaz, I. (2008). Exploring the factors related to acceptance of evolutionary theory among Turkish preservice biology teachers: Toward a more informative conceptual ecology for biological evolution. *Journal of Research in Science Teaching, 45*, 420–443. https://doi.org/10.1002/tea.20223.
Elik, N., Wiener, J., & Corkum, P. (2010). Pre-service teachers' open-minded thinking dispositions, readiness to learn, and attitudes about learning and behavioural difficulties in students. *European Journal of Teacher Education, 33*, 127–146. https://doi.org/10.1080/02619760903524658.
Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology, 71*(2), 390–405. https://doi.org/10.1037/0022-3514.71.2.390.
Flora, D. B., LaBrish, C., & Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2012.00055.
Gerber, S., & Scott, L. (2011). Gamers and gaming context: Relationships to critical thinking: Gamers and critical thinking. *British Journal of Educational Technology, 42*, 842–849. https://doi.org/10.1111/j.1467-8535.2010.01106.x.
Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making, 8*, 188–201.
Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2015). Unraveling the effects of critical thinking instructions, practice, and self-explanation on students' reasoning performance. *Instructional Science, 43*, 487–506. https://doi.org/10.1007/s11251-015-9347-8.
Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2014). Improving critical thinking: Effects of dispositions and instructions on economics students' reasoning skills. *Learning and Instruction, 29*, 31–42. https://doi.org/10.1016/j.learninstruc.2013.07.003.
Heijltjes, A., Van Gog, T., & Paas, F. (2014). Improving students' critical thinking: Empirical support for explicit instructions combined with practice. *Applied Cognitive Psychology, 28*, 518–530. https://doi.org/10.1002/acp.3025.
Hu, L., & 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*, 1–55. https://doi.org/10.1080/10705519909540118.
Jamieson, S. (2004). Likert scales: How to (ab) use them. *Medical Education, 38*, 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x.
Jurkovič, M. (2016). Effect of short-term mindfulness induction on myside bias and miserly processing: A preliminary study. *Studia Psychologica, 58*, 231–237. https://doi.org/10.21909/sp.2016.03.719.
Klaczynski, P. A. (2014). Heuristics and biases: Interactions among numeracy, ability, and reflectiveness predict normative responding. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00665.
Krumrei-Mancuso, E. J., & Rouse, S. V. (2016). The development and validation of the Comprehensive Intellectual Humility Scale. *Journal of Personality Assessment, 98*, 209–221. https://doi.org/10.1080/00223891.2015.1068174.
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*, 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*, 223–248. https://doi.org/10.1080/15305058.2012.703734.
Kuijpers, R. E. (2015). *Applications of categorical marginal models in test construction.* Tilburg University.
Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology, 43*, 42–69. https://doi.org/10.1177/0081175013481958.
Lean Keng, S., & AlQudah, H. N. I. (2017). Assessment of cognitive bias in decision-making and leadership styles among critical care nurses: A mixed methods study. *Journal of Advanced Nursing, 73*(2), 465–481. https://doi.org/10.1111/jan.13142.
Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009.
Magidson, J., & Vermunt, J. K. (2003). Comparing latent class factor analysis with the traditional approach in datamining. In H. Bozdogan (Ed.). *Statistical data mining and knowledge discovery* (pp. 373–383). Chapman & Hall/CRC.
Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*, 293–299. https://doi.org/10.1037/1082-989X.1.3.293.
Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., et al. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology Applied, 21*, 1–14. https://doi.org/10.1037/xap0000040.
Mokken, R. J. (1971). *A theory and procedure of scale analysis.* De Gruyter Mouton.
Molenaar, I. W., & Sijtsma, K. (2000). *MPS5 for Windows. A program for Mokken scale analysis for polytomous items*.
Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition, 42*, 1–10. https://doi.org/10.3758/s13421-013-0340-7.
Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology, 80*, 34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001.
Price, E., Ottati, V., Wilson, C., & Kim, S. (2015). Open-minded cognition. *Personality & Social Psychology Bulletin, 41*, 1488–1504. https://doi.org/10.1177/0146167215600528.
R Development Core Team (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computinghttp://www.R-project.org.
Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553.
Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of Statistical Software, 48*. https://doi.org/10.18637/jss.v048.i02.
Sá, W. C., & Stanovich, K. E. (2001). The domain specificity and generality of mental contamination: Accuracy and projection in judgments of mental content. *British Journal of Psychology, 92*, 281–302. https://doi.org/10.1348/000712601162194.
Sá, W. C., Kelley, C. N., Ho, C., & Stanovich, K. E. (2005). Thinking about personal theories: Individual differences in the coordination of theory and evidence. *Personality and Individual Differences, 38*, 1149–1161. https://doi.org/10.1016/j.paid.2004.07.012.
Sá, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 91*, 497–510. https://doi.org/10.1037/0022-0663.91.3.497.
Shu, S. B., & Townsend, C. (2014). Using aesthetics and self-affirmation to encourage openness to risky (and safe) choices. *Journal of Experimental Psychology Applied, 20*, 22–39. https://doi.org/10.1037/xap0000003.
Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* SAGE.
Smits, I. A. M., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory Mokken scale analysis as a dimensionality assessment tool: Why scalability does not imply unidimensionality. *Applied Psychological Measurement, 36*(6), 516–539. https://doi.org/10.1177/0146621612451050.
Spadaccini, J., & Esteves, J. E. (2014). Intuition, analysis and reflection: An experimental study into the decision-making processes and thinking dispositions of osteopathy students. *International Journal of Osteopathic Medicine, 17*, 263–271. https://doi.org/10.1016/j.ijosm.2014.04.004.
Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*, 342–357. https://doi.org/10.1037/0022-0663.89.2.342.
Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning, 13*, 225–247. https://doi.org/10.1080/13546780600780796.
Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*, 672–695. https://doi.org/10.1037/0022-3514.94.4.672.
Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking.* MIT Press.
Strathman, A., Gleicher, F., Boninger, D. S., & Edwards, C. S. (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology, 66*, 742–752. https://doi.org/10.1037/0022-3514.66.4.742.
Svedholm-Häkkinen, A. M., & Lindeman, M. (2018). Actively open-minded thinking: Development of a shortened scale and disentangling attitudes towards knowledge

and people. *Thinking & Reasoning, 24*, 21–40. https://doi.org/10.1080/13546783.2017.1378723.
Swami, V., Voracek, M., Stieger, S., Tran, U. S., & Furnham, A. (2014). Analytic thinking reduces belief in conspiracy theories. *Cognition, 133*, 572–585. https://doi.org/10.1016/j.cognition.2014.08.006.
Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning, 20*, 215–244. https://doi.org/10.1080/13546783.2013.869763.
Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. S.t. B. T. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology General, 147*, 945–961. https://doi.org/10.1037/xge0000457.
Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., et al. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition, 128*, 237–251. https://doi.org/10.1016/j.cognition.2012.09.012.
Toplak, M. E., & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 94*, 197–209. https://doi.org/10.1037//0022-0663.94.1.197.
Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*, 1275–1289. https://doi.org/10.3758/s13421-011-0104-1.
Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample: Heuristics and biases tasks and outcomes. *Journal of Behavioral Decision Making, 30*, 541–554. https://doi.org/10.1002/bdm.1973.
Toplak, M. E., West, R. F., & Stanovich, K. E. (2014a). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental Psychology, 50*, 1037–1048. https://doi.org/10.1037/a0034910.
Toplak, M. E., West, R. F., & Stanovich, K. E. (2014b). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20*, 147–168. https://doi.org/10.1080/13546783.2013.844729.
Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*, 1–19.
Van der Ark, L. A., Van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement, 35*, 380–392. https://doi.org/10.1177/0146621610392911.
Van der Palm, D. W., Van der Ark, L. A., & Sijtsma, K. (2014). A flexible latent class approach to estimating test-score reliability. *Journal of Educational Measurement, 51*, 339–357. https://doi.org/10.1111/jedm.12053.
West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology, 103*, 506–519. https://doi.org/10.1037/a0028857.
West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology, 100*, 930–941. https://doi.org/10.1037/a0012842.