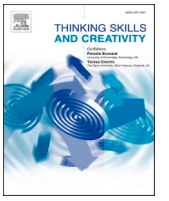
Thinking Skills and Creativity 39 (2021) 100793

Contents lists available at ScienceDirect

Thinking Skills and Creativity

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



Psychometric properties of the short-form CART: Investigating its dimensionality through a Mokken Scale analysis

Marion Tillema [a,b,*], Samantha Bouwmeester [c], Peter Verkoeijen [b,c], Anita Heijltjes [b]

[a] Academy for Industry and Information Sciences, Avans University of Applied Sciences, the Netherlands
[b] Learning and Innovation Centre, Avans University of Applied Sciences, the Netherlands
[c] Department of Psychology and Child Studies, Erasmus School of Social and Behavioural Sciences, Erasmus University Rotterdam, the Netherlands

ARTICLE INFO

Keywords:
Critical thinking
Rationality
Mokken scale analysis
Dimensionality

ABSTRACT

Rationality is considered an important trait for functioning in modern society, as it is thought to be necessary for making judgments and decisions. Rationality aligns well with ambitions in higher education for teaching students to be critical thinkers and decision makers. The short-form Comprehensive Assessment of Rational Thinking (CART) might therefore be suitable for measuring this ability in higher education. Subtest total scores are used and interpreted as distinguishable units of measurement, representing single subconstructs within the concept of rationality. In this study, we explored whether the subtests can indeed be interpreted as unidimensional scales. We assessed the subtests' dimensionality by performing a Mokken scale analysis using the performance data of 185 students of higher education on a Dutch translation of the short-form CART. For two subtests, all items formed a single ordinal scale. For three subtests, a subset of items formed a scale. For seven subtests, no items, nor any subset of items, formed a unidimensional scale. In its current form, then, not all short-form CART subtest scores could be meaningfully interpreted as measures of constituent parts of rationality, limiting its use for assessment in higher education.

1. Introduction

The importance of critical thinking skills as a learning outcome has been advocated by scholars, administrators, and educators (Bok, 2006; Facione, 1990; Halpern, 1999; OECD, 2018; Siegel, 1989). The Organisation for Economic Co-operation and Development (OECD), for example, emphasises that: 'There is a growing consensus that formal education should cultivate the creativity and critical thinking skills of students to help them succeed in modern, globalised economies based on knowledge and innovation' (OECD, 2018). For higher educated professionals, critical thinking skills are often part and parcel of their professional practice (e.g., Van Peppen et al., 2018). Koehler, Brenner, and Griffin (2002), for example, shows how making sound judgments and decisions based on probabilities is essential in performing jobs in medicine, meteorology, law, business, and sports (see also Croskerry, 2009; Rachlinski, 2004).

Given this importance of critical thinking skills in complex dynamical professional environments, it is important to develop teaching programs which foster students' critical thinking skills effectively and efficiently. Valid measurements of critical thinking for conducting both formative and summative assessments are essential to such programs. Only if critical thinking skills are measured validly, can we can we monitor the effectiveness of teaching programs and can we consider obtained diplomas as indicators of

* Corresponding author at: Avans University of Applied Sciences, PO Box 90.116, 4800 RA, Breda, the Netherlands.
E-mail address: m.tillema@avans.nl (M. Tillema).

https://doi.org/10.1016/j.tsc.2021.100793
Received 31 July 2020; Received in revised form 27 January 2021; Accepted 30 January 2021
Available online 6 February 2021

sufficient critical thinking skills (as part of and in addition to professional knowledge and skills). There are a number of standardised instruments published for measuring critical thinking skills generically (e.g. Facione & Facione, 1992; Halpern, 2010; Parker & Fischhoff, 2005). (See Rear (2019) for a discussion of the limitations of generic tests of critical thinking.)

A fairly recent addition to the set of standardised generic instruments was put forth by Stanovich, West, and Toplak (2017) in the form of the Comprehensive Assessment of Rational Thinking, or CART. The CART might be a suitable candidate for measuring critical thinking in higher education[1] . Stanovich et al. (2017) use rationality as a characteristic on which individuals can be ordered relative to each other: responses can be less or more rational in that individuals will vary in the degree to which they 'deviate from the optimum response pattern' for maximizing utility (p.8). Individual ordering makes an instrument suitable for assessing rationality on the individual student level, both formatively and summatively. The authors distinguish between epistemic and instrumental rationality. The first is, to put it colloquially, about 'what is true'. Or, to put it more formally, 'how well our beliefs map onto the actual structure of the world' (p.6). Instrumental rationality is about 'what to do': choosing 'among options based on […] the greatest expected utility' (p.7). Instrumental rationality fits in well with ambitions in higher education for teaching students to be critical thinkers and decision makers.

The theoretical underpinnings of the CART are strongly tied in with the notion of Type II processing as posited in the dual process theory (Croskerry, 2009; Evans, 2011; Kahneman, 2011; Evans & Stanovich, 2013; Wason & Evans, 1974). This model distinguishes a Type I mode of thinking and a Type II[2] mode of thinking. Type I thinking is claimed to be 'fast, high capacity, independent of working memory and cognitive ability' (Evans, 2011, p.87). Stanovich et al. (2017) emphasise that the defining characteristic of Type I thinking is its autonomy: these 'processes execute automatically upon encountering their triggering stimuli' (Stanovich et al., 2017, p. 17). Examples are emotional regulation and implicit learning (Stanovich et al., 2017) and instinctive behaviors and associative learning (Evans, 2003). The Type II mode of thinking is 'slow, low capacity, heavily dependent on working memory and related to individual differences in cognitive ability' (Evans, 2011, p.87). Its defining characteristic is that, contrary to Type I thinking, it 'permits abstract hypothetical thinking' (Evans, 2003, p.454, see also Stanovich et al., 2017, p.323). Type II thinking is thought to be necessary for making judgments (Stanovich et al., 2017; Tversky & Kahneman, 1974) and decisions (Evans, 2011; Stanovich et al., 2017).

Stanovich et al. (2017) postulate that rationality tests, or situations that require rationality to render optimum responses, require Type II thinking to override Type I thinking, but contain no salient warnings that Type II thinking is necessary: the test taker must make the effort to activate Type II thinking without receiving cues to do so. The process of activating Type II thinking to arrive at an instrumentally rational response involves the following three steps, in which there is interplay between cognitive processing capacities and declarative and procedural knowledge structures. The first step is detection of an alternative to the intuitive response. For this, a respondent needs relevant knowledge structures. For example: when encountered with a patient who was admitted to the hospital with anxiety and now shows symptoms of chest pain, a medical professional will need to have declarative knowledge of all possible diagnoses explaining chest pain in order to signal that there are alternatives to the intuitive response – attributing chest pain to anxiety – such as respiratory and heart conditions (Croskerry, Singhal, & Mamede, 2013). The next step involves processing capacities: sustaining the override[3] of the intuitive response to engage in more effortful and deliberate Type II thinking and decision making processes, followed by the step of applying declarative knowledge and procedural knowledge (rules and strategies) to arrive at the optimum response (see also Bonnefon, 2018). In the case of this example the latter would involve engaging in the procedure of methodical diagnosis testing. The knowledge structures in this process are also referred to as mindware by Stanovich et al. (2017). In addition to helpful mindware, the authors also identify contaminated mindware, such as dysfunctional personal beliefs which foster irrational thinking. Problems arise if helpful mindware or knowledge is absent (e.g. Type II processing is not activated due to 'a mindware gap') or if contaminated mindware is present.

The CART consists of 24 subtests which are designed to tap this interplay of declarative and procedural knowledge and processing capacities, which forms Type II thinking. The Reflection vs Intuition subtests, for example, is designed to mainly draw on detection and override capacities, whereas the Scientific Reasoning subtest should rely on override processing as well as declarative knowledge. In addition, the test battery also includes four thinking dispositions tests. Dispositions are, by the authors, not considered part of rationality itself, but a prerequisite for rationality. There is also a short-form CART, which is limited to twelve subtests, covering the same four aspects processing, knowledge (helpful mindware), contaminated mindware and thinking dispositions. The full-form CART contains six more subtests than the short-form version in processing, three more knowledge subtests and three more thinking dispositions subtests. No clear rationale is offered for the selection of subtests in the short-form CART. However, the short-form CART is presented as a legitimate test in itself, with the advantage of logistic feasibility in educational practice: it takes about two hours to complete. Table 1 shows an overview of the subtests in the short-form CART as categorised by the authors under one of the four aspects. Only the Probabilistic and Statistical Reasoning subtests and the Scientific Reasoning subtest are listed under two aspects.

It is important that tests of critical thinking such as the short-form CART are valid measurements. A test's internal structure provides important information regarding its validity (Leach, Immekus, French, & Hand, 2020). The CART's theoretical underpinnings suggest a multidimensional conceptualization of rationality, involving the aforementioned processing capacities and mindware

[1] Ideally, the inclusion of domain specific elements would seem advisable in addition to the generic operationalization (see Rear, 2018). This, however, is not the focus of this study.

[2] Some researchers (e.g. Croskerry, 2009; Kahneman, 2011) have used the terms System I and System II instead of Type I and Type II. Evans (2011), however, explains why 'Type' seems to be a more fitting label than 'System'.

[3] In some cases the patterns of Type II thinking will be so automated, for example due to practice and experience, that Type I will render the optimum response, in this case: testing alternative diagnoses, leaving no need for engaging in Type II thinking.

Table 1
Subtests in the short-form CART, adapted from: Stanovich, West & Toplak (2017, p. 221).

Short-form CART			Thinking dispositions
Processing	Knowledge	Contaminated mindware	
Probabilistic and statistical reasoning		Superstitious thinking	Actively Open-Minded Thinking Scale
Scientific reasoning		Anti-science attitudes	
Reflection vs intuition	Probabilistic Numeracy	Conspiracy beliefs	
Syllogisms		Dysfunctional Personal Beliefs	
Ratio bias			
Disjunctive reasoning			

structures. Specific subtests are, on theoretical grounds, assumed to tap into an interplay between mindware and processing, or into processing only. Stanovich et al. (2017) report the results of a Principal Components Analysis on the subtest total scores, rendering three components: one containing seven subtests that were a on theoretical grounds listed under 'processing' and 'knowledge', one containing three of the four subtests listed under 'avoidance of contaminated mindware', and one containing a subtest in 'processing' and a subtest in 'avoidance of contaminated mindware'. In short, the component structure suggests acceptable validity: the components seem in broad agreement with the theoretically posited dimensions.

However, while component scores seem warranted given the outcomes of the Principal Components Analysis, Stanovich et al. (2017) mainly use and interpret subtest scores (besides short-form CART total tests scores). This signals that subtests are considered distinguishable units of measurement, with the subtest total scores representing single subconstructs. (These subconstructs, then, may cluster into components, such as demonstrated with Stanovich et al. (2017)'s Principal Components Analysis results.) Therefore, before any analyses are performed on the subtest totals, the validity of subtests themselves should be inspected. This means it is important to examine if subtests can indeed be interpreted as unidimensionial manifestations of single constructs, warranting subtest sum scores. Such an interpretation is often made if reliability indicators, such as Cronbach's Alpha, have a sufficiently high value. Indeed, Stanovich et al. (2017) report Cronbachs's alpha per subtest[4] . However, while sometimes used as such, Cronbach's alpha does not give insight into the dimensionality of a test. Sijtsma (2009) explains how Cronbach's alpha can, counter to what is common practice in research in psychology, not be interpreted as a measure of internal consistency: alpha can have the same value for uni- or multifactorial tests (i.e. uni- or multidimensional tests). In other words, reliability measures such as Cronbach's alpha, cannot be used for determining whether or not a test is unidimensional.

In the present study, the question whether short-form CART subtests can be interpreted as unidimensionial manifestations of single constructs will be investigated with Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002), an application of item response theory. This analysis is suitable for assessing the short-form CART's subtests' dimensionality and to address the question whether the subtest items load onto a single underlying latent construct. First, it is suitable for dichotomous or polytomous data (Sijtsma & Molenaar, 2002). Furthermore, it assesses the scalability of items on a dimension by maximizing the explained variation that items have in common, as expressed by Loevinger's H coefficient (Bedford et al., 2010). As this dimension is described by the shared variance among the items and not by a function which may only be linear, as is the case with Principal Component Analysis, Mokken scale analysis is more flexible in the type and number of dimensions it accommodates. In addition, Mokken scale analysis does not only explore each item's scalability, but it also gives us information about an item's discriminative power (the H coefficient should exceed a prefixed lowerbound value).

To sum up, this study assesses the fitness for purpose of the short-form CART by investigating the following questions regarding the subtests' validity:

1 Which test items in the short-form CART subtests are scalable, assuming a single underlying latent trait per subtest?
2 Can the sum scores for subtests be used to discriminate between individual test takers' ability on the assumed underlying trait?

2. Method

2.1. Respondents

Respondents were 185 first-year students of Mechanical Engineering ($n = 78$) and Electrical and Computer Engineering ($n = 107$) at [name of large institute for higher education], [name of European country]. Their mean age was 19.2 ($Sd = 1.8$). Although research skills are part of the curriculum in the first and second year (mostly: formulating problem definitions and research questions, and executing pre-structured research methods), the lesson programmes do not include any explicit training in the types of tasks present in the short-form CART.

[4] Values range from .49 to .93, with values below .70 for five of the short-form CART's subtests.

2.2. Materials

The short-form CART holds a mix of response formats, consisting of multiple choice questions, short open answers questions and scale responses (often with six scale points). Maximum and realised maximum scores per subtest can be found in Table 2. The principal researcher of the present study translated all the items from the short-form CART into Dutch, the students' native language and the working language of their university. The researcher was deemed competent to do this because of a propaedeutic diploma in English Language & Literature, an MA in Linguistics, over ten years of working experience with academic English, eight years of experience teaching Dutch language skills in a university of applied sciences, and familiarity with the research body on rationality and the test types in the CART. Care was taken that complexity of wording an grammar and ordering of information in the translated version was similar to the original version. Eleven out of the 24 items in the Conspiracy Beliefs subtest were specific to a North American context. In such cases, it would become likely that such an items would lose all discriminative power, as none of the respondents in this study would be likely to have heard of the contents. These items were therefore adjusted to a comparable Dutch context. An item containing an unjustified belief about concealment of evidence of health risks of fluoridated water, for example, was replaced by a statement about concealment of evidence of unsafe nuclear reactors. Fluoridation of public drinking water does not take place in the Netherlands, and is not the topic of conspiracy theories, or much of a topic of general interest or public debate at all. The alternative item om nuclear reactions is comparable in terms of relative believability, topic (perceived health risks and concealment) and how long this idea has been around. The item of Some examples of translated items can be found in Appendix A.

2.3. Procedures

The students completed the short-form CART in Qualtrics (Qualtrics, Provo, UT) in a classroom setting and in the presence of a test leader to supervise, who was also a teacher in the institute. There were two sessions: one for the students of Mechanical Engineering and one for the students of Electrical and Computer Engineering. The students were seated at group tables of about five persons per table, and worked on their own laptops. The test supervisor had a good overview of the classroom and ensured silence during the session. There was no possibility for students to copy each other's answers. They had 2,5 hours to complete the short-form CART, with the possibility to take a short break in the classroom at a self-chosen moment. The short-form CART items were presented as instructed by the test makers, in terms of item order.

2.4. Analyses

The response data were recoded into items scores according to the test instructions in R 4.0.0 (RStudio Team, 2020). This rendered a dichotomous data set for the seven subtests listed under knowledge and processing and a polytomous data set for the four subtests listed under contaminated mindware. There were missing values. This might be due to the length and duration of the test. Students also reported they found the recurrence of similar items tedious. On average, the percentage missings per item was 20 % ($Sd = 14.9$, min. = 6, max. = 100). The percentage increased as the test progressed. As Mokken scale analysis is not possible with missing values, three steps were taken. First, for each subtest separately, all participants with more than fifty percent missing responses were removed from the data set. Subsequently, for each subtest separately, all items with more than fifty percent missing responses were removed from the data set. The remaining numbers of participants and items for analyses may be found in Table 3. Third, for subtests A through G, i.e. the subtests (assumedly) on knowledge and processing (Table 1), the remaining missing responses were then treated as incorrect answers

Table 2
Possible raw score ranges, Minimum (Min) and Maximum (Max) Scores in this sample, Mean scores (M) and the Corresponding Standard Deviations (SD) and Subtest Total-Scalability Coefficients (H) and the Corresponding Standard Errors for the 12 Short-form CART subtests.

Subtest	Range	Min	Max	M (SD)	H (SE)
A (Probabilistic and Statistical Reasoning)	0–18	0	17	11.25 (7.41)	0.120 (0.020)
B (Scientific Reasoning)	0–20	0	13	6.72 (5.72)	0.047 (0.021)
C (Reflection vs Intuition)	0–10	0	9	4.20 (4.31)	0.312 (0.039)
D (Syllogisms)	0–16	3	15	9.15 (2.71)	0.079 (0.017)
E (Ratio Bias)	0–12[a]	0	12	7.75 (5.65)	0.196 (0.030)
F (Disjunctive Reasoning)	0–6	0	6	2.72 (2.89)	0.394 (0.047)
G (Probabilistic Numeracy)	0–9	0	8	3.79 (3.74)	0.507 (0.042)
H (Superstitious Thinking)	0–72	11	66	45.98 (16.62)	0.373 (0.032)
I (Anti-Science Attitudes)	0–78	13	78	50.21 (17.68)	0.167 (0.031)
J (Conspiracy Beliefs)	0–144	24	144	94.12 (35.70)	0.454 (0.042)
K (Dysfunctional Beliefs)	0–54	9	54	33.26 (12.49)	0.167 (0.030)
L (Actively Open-Minded Thinking)	0–180	30	180	96.94 (42.72)	0.106 (0.020)

[a] The scoring instructions for subtest E render a somewhat ambiguous data type. The answer options are 1–6, with answer options 4, 5 and 6 all marked as correct, though with 6 (strongly prefer small…) being somewhat more correct than 4 (slightly prefer small…), and answer options 1 -3 (strongly prefer large … slightly prefer large…) marked as incorrect. This renders a data type that seems to linger somewhere between dichotomous (there are two classes of answers: correct and incorrect) and ordinal (answer options 1, 2, 3 can be ordered ordinally, and answer options 4, 5, 6 can be ordered ordinally, but the answer set 1–6 cannot). Out of these two options, we have opted for treating the data for subtest E as dichotomous: the difference of meaning between correct (4, 5, 6) and incorrect (1, 2, 3) is most pertinent to the interpretation of the subtest.

Table 3
Number of participants (out of 185) and items (original number of items per subtest between brackets, *: not counting filler items) included in the analyses per subtest after removing participants and items with > 50 % missing values.

Subtest	N participants	N items
A (Probabilistic and Statistical Reasoning)	171	17 (18)
B (Scientific Reasoning)	165	13 (18)
C (Reflection vs Intuition)	157	9 (11)
D (Syllogisms)	155	16 (16)
E (Ratio Bias)	154	12 (12*)
F (Disjunctive Reasoning)	157	6 (6)
G (Probabilistic Numeracy)	150	8 (9)
H (Superstitious Thinking)	148	11 (12)
I (Anti-Science Attitudes)	153	13 (13)
J (Conspiracy Beliefs)	150	24 (24*)
K (Dysfunctional Beliefs)	148	9 (9)
L (Actively Open-Minded Thinking)	148	30 (30)

and therefore coded as 0. For subtests H through L, i.e. the subtests (assumedly) representing a tendency toward avoidance of contaminated mindware or a thinking disposition, the remaining missing responses were imputed with multiple imputation using Fully Conditional Specification (FCS) implemented by the MICE algorithm (Van Buuren & Groothuis-Oudshoorn, 2011).

Mokken scale analysis (package 3.0.4) was run for each subtest separately. A subset of items are assumed to form a Mokken scale when they form a unidimensional scale, when there is local independence conditional on the total score and when people can reliably be ordered according to their scale scores. The scalability of the subset of items is indicated by *H*. It is assumed that *H* values below .3 indicate no scale, values between .3 and .4 indicate a weak scale, values between .4 and .5 indicate a medium scale and *H*'s larger than .6 indicate a strong scale (Sijtsma & Molenaar, 2002). The H_i indicates the scalability of an item and can be used to assess whether an individual item fits into the scale. We used the Automatic Item Selection Procedure from the Mokken package in R (R Development Core Team, 2008; Van der Ark, 2007) to evaluate the scalability of the subtests. A lowerbound value of $H>.3$ was used as the criterion to select items. Zero-variance items were excluded from the analysis beforehand. These were present in three subtests: Scientific Reasoning, Reflection versus Intuition, and Probabilistic Numeracy.

3. Results

Columns 2 through 5 of Table 2 report relevant descriptive statistics about each of the subtests. Performance on all subtests is generally intermediate, with mean scores in the intermediate range of the scale. The maximum achieved scores generally approach the highest possible score, with subtest B (Scientific Reasoning) being a notable exception, with a maximum score of 13 out of 20.

Column 6 of Table 2 shows that two of the twelve subtests had scalability coefficients (*H*) larger than the recommended minimum lowerbound for discriminating between individuals of .4 for total-scale coefficients (Mokken, 1971; Sijtsma & Molenaar, 2002). These were the Probabilistic Numeracy subtest and the Conspiracy Beliefs subtest. As Table 4 shows, all the analysed items in these subtest were scalable onto one scale (per subtest), with scalability coefficients (H_i) above .3 (as recommended for item-scale coefficients, Sijtsma & Molenaar, 2002). Hence, the sum scores for these two tests can be used meaningfully.

Three other subtests approached acceptable scalability (Table 2), with *H* values of .312 (Reflection vs Intuition subtest), .394 (Disjunctive Reasoning subtest) and .373 (Superstitious Thinking subtest). Table 4 shows that in the Reflection vs Intuition subtest, two items were removed from the analysis due to too many missing values, two items were non-scalable, five items clustered onto one scale (but $H_i < .3$ for one item) and two clustered onto another scale (but $H_i < .3$ for one item), leaving us with just four out of eleven items with $H_i > .3$ and clustering on one scale. In the Disjunctive Reasoning subtest, five items clustered onto one scale (but $H_i < .3$ for one item), and one item was non-scalable, leaving us with four out of six items with $H_i > .3$ and clustering onto one scale. In the Superstitious Thinking subtest, one item was removed from the analysis due to too many missing values, nine items clustered onto one scale ($H_i > .3$ for all items) and two items clustered onto another scale ($H_i < .3$ for both items). In the remaining seven subtests ($H < .3$, Table 2), no items discriminated sufficiently ($H_i < .3$, Table 4).

4. Discussion

For two subtests, namely the Probabilistic Numeracy subtest and the Conspiracy Beliefs Subtest, the total set of items and therefore the subtest sum scores can be used for discriminating between students on the (assumed) underlying latent trait. Therefore, the sum scores for these subtests can be interpreted meaningfully. It should be noted, however, that item with too many missing values, which was also a zero-variance, was excluded from the analysis for Probabilistic Numeracy, which of course indicates that this item did not discriminate. For three subtests, a subselection of items with acceptable discriminative power might be summed into a score reflecting a single trait, although the number of remaining items is rather low: four out of eleven items for Reflection vs Intuition, four out of six items for Disjunctive Reasoning, and nine out of twelve items for Superstitious Thinking. For seven subtests, no total or subset sum scores can be used meaningfully to reflect a single trait.

In Stanovich et al. (2017) the short-form CART was examined by administering it to a volunteer group of university students with a mean age of 18.7. There are some differences between the mean scores and standard deviations in the present study and Stanovich

Table 4
Scales, Mean scores (M) and the Corresponding Standard Deviations (SD) and Item Scalability Coefficients (Hj) for all items.

Item	Scale	Range	Min	Max	M (SD)	H_i (SE)
Subtest A (Probabilistic and Statistical Reasoning)						
1	1	0–1	0	1	0.57 (0.50)	0.205 (0.034)
2	1	0–1	0	1	0.46 (0.50)	0.268 (0.033)
3	1	0–1	0	1	0.61 (0.49)	0.162 (0.035)
4	removed, > 50 % missing					
5	1	0–1	0	1	0.86 (0.35)	0.139 (0.058)
6	1	0–1	0	1	0.81 (0.39)	0.177 (0.044)
7	0	0–1	0	1	0.83 (0.38)	0.137 (0.056)
8	1	0–1	0	1	0.87 (0.34)	0.285 (0.043)
9	1	0–1	0	1	0.78 (0.41)	0.220 (0.039)
10	–	0–1	0	1	0.56 (0.50)	0.090 (0.039)
11	–	0–1	0	1	0.42 (0.49)	0.015 (0.048)
12	–	0–1	0	1	0.44 (0.50)	−0.027 (0.046)
13	–	0–1	0	1	0.79 (0.41)	0.001 (0.044)
14	–	0–1	0	1	0.61 (0.49)	0.067 (0.039)
15	–	0–1	0	1	0.95 (0.21)	0.247 (0.089)
16	–	0–1	0	1	0.46 (0.50)	0.010 (0.043)
17	2	0–1	0	1	0.64 (0.48)	0.086 (0.039)
18	2	0–1	0	1	0.69 (0.46)	0.136 (0.038)
Subtest B (Scientific Reasoning)						
1	removed, > 50 % missing					
2	removed, > 50 % missing					
3	removed, > 50 % missing					
4	removed, > 50 % missing					
5	–	0–1	0	1	0.10 (0.30)	−0.082 (0.068)
6	–	0–1	0	1	0.61 (0.49)	0.033 (0.038)
7	–	0–1	0	1	0.87 (0.34)	0.030 (0.065)
8	1	0–1	0	1	0.66 (0.47)	0.052 (0.041)
9	1	0–1	0	1	0.52 (0.50)	0.019 (0.039)
10	–	0–1	0	1	0.26 (0.44)	0.026 (0.045)
11	–	0–1	0	1	0.47 (0.50)	0.082 (0.038)
12	–	0–1	0	1	0.39 (0.49)	0.069 (0.039)
13	2	0–1	0	1	0.29 (0.45)	0.046 (0.042)
14	–	0–1	0	1	0.60 (0.49)	0.061 (0.037)
15	2	0–1	0	1	0.35 (0.48)	0.095 (0.040)
16	–	0–1	0	1	0.75 (0.43)	0.023 (0.046)
17	–	0–1	0	1	0.87 (0.34)	0.068 (0.056)
18	removed, > 50 % missing					
Subtest C (Reflection vs Intuition)						
1	–	0–1	0	1	0.55 (0.50)	0.256 (0.053)
2	2	0–1	0	1	0.41 (0.49)	0.277 (0.057)
3	1	0–1	0	1	0.58 (0.49)	0.339 (0.048)
4	2	0–1	0	1	0.64 (0.48)	0.305 (0.060)
5	1	0–1	0	1	0.60 (0.49)	0.379 (0.048)
6	–	0–1	0	1	0.33 (0.47)	0.259 (0.069)
7	removed, > 50 % missing					
8	removed, > 50 % missing					
9	1	0–1	0	1	0.50 (0.50)	0.238 (0.055)
10	1	0–1	0	1	0.40 (0.49)	0.395 (0.053)
11	1	0–1	0	1	0.19 (0.39)	0.399 (0.086)
Subtest D (Syllogisms)						
1	1	0–1	0	1	0.70 (0.46)	0.050 (0.034)
2	–	0–1	0	1	0.62 (0.49)	0.105 (0.031)
3	–	0–1	0	1	0.57 (0.50)	0.094 (0.033)
4	2	0–1	0	1	0.47 (0.50)	0.132 (0.035)
5	–	0–1	0	1	0.38 (0.49)	−0.153 (0.045)
6	1	0–1	0	1	0.59 (0.49)	0.120 (0.031)
7	2	0–1	0	1	0.45 (0.50)	0.114 (0.037)
8	–	0–1	0	1	0.68 (0.47)	0.094 (0.032)
9	–	0–1	0	1	0.63 (0.48)	0.053 (0.034)
10	–	0–1	0	1	0.50 (0.50)	0.039 (0.036)
11	0	0–1	0	1	0.68 (0.47)	0.128 (0.030)
12	2	0–1	0	1	0.63 (0.48)	0.114 (0.031)
13	–	0–1	0	1	0.50 (0.50)	0.041 (0.036)
14	–	0–1	0	1	0.64 (0.48)	0.106 (0.030)
15	2	0–1	0	1	0.47 (0.50)	0.068 (0.037)
16	–	0–1	0	1	0.66 (0.47)	0.117 (0.031)
Subtest E (Ratio Bias)						

(*continued on next page*)

Table 4 (*continued*)

Item	Scale	Range	Min	Max	M (SD)	H_i (SE)
1	–	0–1	0	1	0.77 (0.42)	0.198 (0.049)
2	filler item					
3	1	0–1	0	1	0.77 (0.42)	0.235 (0.043)
4	2	0–1	0	1	0.66 (0.48)	0.131 (0.043)
5	–	0–1	0	1	0.63 (0.48)	0.214 (0.042)
6	filler item					
7	–	0–1	0	1	0.7 (0.46)	0.192 (0.044)
8	–	0–1	0	1	0.66 (0.47)	0.213 (0.040)
9	–	0–1	0	1	0.68 (0.47)	0.172 (0.044)
10	filler item					
11	–	0–1	0	1	0.60 (0.49)	0.181 (0.044)
12	–	0–1	0	1	0.57 (0.50)	0.177 (0.049)
13	1	0–1	0	1	0.58 (0.49)	0.232 (0.044)
14	2	0–1	0	1	0.49 (0.5)	0.205 (0.056)
15	1	0–1	0	1	0.65 (0.48)	0.217 (0.041)
Subtest F (Disjunctive Reasoning)						
1	1	0–1	0	1	0.27 (0.44)	0.280 (0.083)
2	1	0–1	0	1	0.58 (0.49)	0.469 (0.056)
3	1	0–1	0	1	0.55 (0.50)	0.397 (0.058)
4	1	0–1	0	1	0.52 (0.50)	0.450 (0.053)
5	1	0–1	0	1	0.50 (0.50)	0.465 (0.054)
6	–	0–1	0	1	0.30 (0.46)	0.243 (0.080)
Subtest G (Probabilistic Numeracy)						
1	1	0–1	0	1	0.69 (0.46)	0.449 (0.070)
2	1	0–1	0	1	0.70 (0.46)	0.540 (0.068)
3	removed, > 50 % missing					
4	1	0–1	0	1	0.27 (0.45)	0.528 (0.057)
5	1	0–1	0	1	0.55 (0.50)	0.528 (0.057)
6	1	0–1	0	1	0.33 (0.60)	0.422 (0.070)
7	1	0–1	0	1	0.60 (0.49)	0.526 (0.057)
8	1	0–1	0	1	0.24 (0.43)	0.452 (0.076)
9	1	0–1	0	1	0.41 (0.49)	0.566 (0.052)
Subtest H (Superstitious Thinking)						
1	1	1–6	1	6	4.13 (1.55)	0.466 (0.038)
2	1	1–6	1	6	4.32 (1.50)	0.478 (0.037)
3	2	1–6	1	6	4.04 (1.50)	0.164 (0.063)
4	removed, > 50 % missing					
5	1	1–6	1	6	4.29 (1.57)	0.409 (0.042)
6	2	1–6	1	6	3.78 (1.54)	0.217 (0.063)
7	1	1–6	1	6	3.76 (1.51)	0.354 (0.051)
8	1	1–6	1	6	4.54 (1.49)	0.434 (0.043)
9	1	1–6	1	6	4.09 (1.49)	0.368 (0.048)
10	1	1–6	1	6	4.26 (1.43)	0.376 (0.050)
11	1	1–6	1	6	4.50 (1.51)	0.467 (0.033)
12	1	1–6	1	6	4.26 (1.52)	0.366 (0.052)
Subtest I (Anti-Science Attitudes)						
1	1	1–6	1	6	3.54 (1.51)	0.156 (0.043)
2	2	1–6	1	6	4.06 (1.23)	0.195 (0.040)
3	1	1–6	1	6	3.93 (1.35)	0.217 (0.042)
4	4	1–6	1	6	3.61 (1.28)	0.083 (0.051)
5	3	1–6	1	6	3.70 (1.47)	0.189 (0.040)
6	3	1–6	1	6	3.83 (1.46)	0.177 (0.041)
7	1	1–6	1	6	4.05 (1.40)	0.195 (0.043)
8	1	1–6	1	6	3.91 (1.22)	0.196 (0.047)
9	4	1–6	1	6	3.88 (1.35)	0.158 (0.045)
10	2	1–6	1	6	3.87 (1.33)	0.174 (0.043)
11	2	1–6	1	6	3.81 (1.29)	0.039 (0.039)
12	2	1–6	1	6	3.95 (1.41)	0.199 (0.047)
13	1	1–6	1	6	4.08 (1.37)	0.181 (0.044)
Subtest J (Conspiracy Beliefs)						
1		1–6	1	6	4.02 (1.54)	0.368 (0.054)
2		1–6	1	6	4.07 (1.31)	0.443 (0.055)
3		1–6	1	6	3.77 (1.31)	0.484 (0.046)
4	justified belief item, not scored					
5		1–6	1	6	3.75 (1.54)	0.438 (0.052)
6		1–6	1	6	3.73 (1.64)	0.405 (0.048)
7	justified belief item, not scored					
8	justified belief item, not scored					
9		1–6	1	6	3.89 (1.40)	0.481 (0.046)

(*continued on next page*)

Table 4 (*continued*)

Item	Scale	Range	Min	Max	M (SD)	H_i (SE)
10		1–6	1	6	4.05 (1.51)	0.420 (0.050)
11		1–6	1	6	4.05 (1.54)	0.420 (0.054)
12		1–6	1	6	4.19 (1.51)	0.481 (0.045)
13		1–6	1	6	4.11 (1.52)	0.470 (0.047)
14		1–6	1	6	3.84 (1.45)	0.428 (0.053)
15		1–6	1	6	4.01 (1.61)	0.462 (0.044)
16		1–6	1	6	3.86 (1.51)	0.448 (0.048)
17		1–6	1	6	3.77 (1.44)	0.508 (0.046)
18	justified belief item, not scored					
19		1–6	1	6	3.75 (1.46)	0.355 (0.061)
20		1–6	1	6	4.03 (1.48)	0.477 (0.048)
21		1–6	1	6	4.03 (1.51)	0.468 (0.047)
22		1–6	1	6	3.85 (1.36)	0.442 (0.056)
23		1–6	1	6	3.78 (1.41)	0.489 (0.050)
24		1–6	1	6	3.88 (1.40)	0.490 (0.050)
25		1–6	1	6	4.02 (1.58)	0.492 (0.043)
26		1–6	1	6	3.91 (1.64)	0.433 (0.054)
27		1–6	1	6	3.89 (1.57)	0.508 (0.044)
28	justified belief item, not scored					
29		1–6	1	6	3.87 (1.50)	0.504 (0.045)
Subtest K (Dysfunctional Beliefs)						
1	1	1–6	1	6	3.84 (1.43)	0.264 (0.042)
2	–	1–6	1	6	3.46 (1.31)	0.149 (0.052)
3	1	1–6	1	6	3.76 (1.43)	0.214 (0.045)
4	2	1–6	1	6	3.71 (1.30)	0.041 (0.062)
5	1	1–6	1	6	3.78 (1.66)	0.262 (0.037)
6	–	1–6	1	6	3.39 (1.31)	−0.149 (0.067)
7	–	1–6	1	6	3.76 (1.30)	0.176 (0.052)
8	1	1–6	1	6	3.86 (1.38)	0.266 (0.040)
9	2	1–6	1	6	3.70 (1.37)	0.236 (0.043)
Subtest L (Actively Open-Minded Thinking)						
1	1	1–6	1	6	2.89 (1.48)	0.216 (0.026)
2	1	1–6	1	6	3.36 (1.34)	0.104 (0.033)
3	1	1–6	1	6	3.11 (1.47)	0.114 (0.034)
4	2	1–6	1	6	3.25 (1.45)	0.180 (0.031)
5	3	1–6	1	6	2.91 (1.43)	0.055 (0.033)
6	4	1–6	1	6	3.22 (1.42)	0.061 (0.038)
7	1	1–6	1	6	2.86 (1.44)	0.154 (0.030)
8	2	1–6	1	6	3.26 (1.36)	0.159 (0.032)
9	–	1–6	1	6	3.44 (1.47)	0.080 (0.042)
10	1	1–6	1	6	3.23 (1.37)	0.060 (0.041)
11	4	1–6	1	6	3.14 (1.62)	0.089 (0.030)
12	–	1–6	1	6	3.10 (1.37)	0.104 (0.040)
13	1	1–6	1	6	3.16 (1.39)	0.138 (0.034)
14	2	1–6	1	6	3.48 (1.45)	0.152 (0.037)
15	2	1–6	1	6	3.33 (1.51)	0.044 (0.040)
16	3	1–6	1	6	3.14 (1.30)	0.066 (0.037)
17	2	1–6	1	6	3.20 (1.44)	0.148 (0.032)
18	–	1–6	1	6	3.35 (1.40)	0.085 (0.037)
19	–	1–6	1	6	3.45 (1.41)	0.079 (0.040)
20	2	1–6	1	6	3.16 (1.44)	0.064 (0.039)
21	1	1–6	1	6	3.29 (1.36)	0.135 (0.032)
22	2	1–6	1	6	3.29 (1.41)	0.124 (0.036)
23	–	1–6	1	6	3.65 (1.44)	−0.060 (0.041)
24	1	1–6	1	6	3.16 (1.40)	0.139 (0.036)
25	1	1–6	1	6	3.04 (1.43)	0.149 (0.031)
26	2	1–6	1	6	3.24 (1.43)	0.131 (0.035)
27	1	1–6	1	6	3.17 (1.50)	0.161 (0.034)
28	–	1–6	1	6	3.66 (1.48)	0.052 (0.041)
29	2	1–6	1	6	3.25 (1.40)	0.093 (0.037)
30	1	1–6	1	6	3.16 (1.32)	0.103 (0.035)

et al.'s (2017) results (see Table 5), but all in all, the participants in this study seem to fall into the intended population. Stanovich et al. (2017) report Cronbach's alpha's for each of the subtests, with no ceiling or floor effects in performance. As was described in the above, however, Cronbach's alpha cannot be used for determining whether or not a test is unidimensional, as it may have the same value for uni- or multidimensional tests (Sijtsma, 2009): it does not express whether or not a set of items taps into the same underlying latent trait. To our knowledge, this is the first study that investigated the psychometric properties of the (short-form) CART using Mokken scale analysis, which is suitable for assessing the short-form CART's dimensionality. For one subtest, to our knowledge, these results

Table 5
Minimum (Min) and Maximum (Max) Scores in sample, Mean scores (M) and the Corresponding Standard Deviations (SD) for the present study and as reported in Stanovich et al. (2017).

	Present study		Stanovich et al. (2017)	
Subtest	Min-Max	M (SD)	Min-Max	M (SD)
A (Probabilistic and Statistical Reasoning)	0–17	11.25 (7.41)	3–17	9.84 (2.84)
B (Scientific Reasoning)	0–13	6.72 (5.72)	2–17	9.07 (3.03)
C (Reflection vs Intuition)	0–9	4.20 (4.31)	0–10	2.40 (2.13)
D (Syllogisms)	3–15	9.15 (2.71)	4–16	9.85 (2.37)
E (Ratio Bias)[a]	0–12	7.75 (5.65)	12–72	46.40 (11.35)
F (Disjunctive Reasoning)	0–6	2.72 (2.89)	0–6	3.45 (1.74)
G (Probabilistic Numeracy)	0–8	3.79 (3.74)	0–9	4.43 (1.58)
H (Superstitious Thinking)	11–66	45.98 (16.62)	12–58	33.17 (9.73)
I (Anti-Science Attitudes)	13–78	50.21 (17.68)	13–58	40.72 (7.55)
J (Conspiracy Beliefs)	24–144	94.12 (35.70)	24–111	63.08 (18.60)
K (Dysfunctional Beliefs)	9–54	33.26 (12.49)	13–53	32.27 (6.59)
L (Actively Open-Minded Thinking)	30–180	96.94 (42.72)	83–163	121.70 (12.32)

[a] We treated these items as dichotomous, see footnote 5, rendering a possible score range of 0–12 instead of 0–72.

corroborate previous research also using Mokken scale analysis: Janssen et al. (2020) performed a Mokken scale analysis on a Dutch translation of the Actively Open-Minded Thinking Scale and also found that none of the items had sufficient discriminative power, as is the case in the current study. As there are no scaling items at all, it is hard to identify a starting point for improving the test: there are no clues as to which type of item content (e.g. pertaining to beliefs or to actions) seem to cluster. A suggestion for further research could be to perform the same analysis on the original Actively Open-Minded Thinking Scale in English. Furthermore, one might wonder if reporting on a disposition (e.g. 'it is important to persevere in your beliefs even when evidence is brought to bear against them') reflects and individual's tendency to display this disposition in actual decisions and actions.

One of the advantages of applying an IRT model for evaluating scalability at the item level, is the possibility to select items, which have added value in terms of discriminative power and to leave out items for which this is lacking. As the short-form CART is in its current form still a fairly lengthy test – it takes at least two hours to complete – this could enhance the test's practical usability. In this case, however, leaving out ill-functioning items would lead to cutting seven subtests: more than half of the short-form CART, which seems at odds with its theoretical underpinnings. There would, for instance, be no subtests left which are assumed to tap into processing or thinking dispositions. A first step would therefore be to improve the subtests. However, the lack of scalability for all items in these subtests give us little indication of what kind of alterations could be made to improve their scalability. If we assume that the traits that these subtests are intended to measure do indeed exist, for which there are theoretical grounds (see e.g. Evans, 2011; Klaczynski & Felmban, 2020 and Stanovich et al., 2017), then possibly a different operationalization is needed. This might involve changes at the item level, e.g. different content.

For example, the validity of subtests assumed to tap into the processing aspect of rationality, such as the Ratio Bias or Disjunctive Reasoning subtest, might be improved if scoring is not only based on the degree to which optimum responses are achieved. Stanovich et al. (2017, p. 132) explain how, for instance, the phenomenon of ratio bias illustrates the failure of sustained override of Type I processing. However, the presence or absence of an optimum response is an indirect measure, which cannot inform us unequivocally whether override has occurred or not. Indeed, Stanovich et al. (2017) themselves explain how optimum responses might be reached by heuristic responses, which do not necessitate occurrence of the override. Possibly, the processing subtests could provide richer information, and as such possibly become more valid, by combining response outcomes (correct answers or decisions or lack thereof) with process measures. Frey, Johnson, and De Neys (2018) forward some process measures to reflect the conflict detection which is prerequisite for override to occur, such as combined confidence ratings and response times. Yet, it should be noted that these measures might lead to validity issues, in that confidence ratings and response times seem to reflect different parts of the conflict detection construct and in terms of how well a response times as a scale measurement can reflect the binary construct that is detection. In addition, Stanovich et al. (2017, p.56) explain that sustained override involves various stages: successfully carried out override involves, with the prerequisite if having the required mindware, detecting an alternative response, sustained inhibition of Type I processing, and resourcing to sufficient cognitive capacity. In other words: it could go wrong in various stages. For this reason, a dynamic assessment (see e.g. Fabio, 2005) might be suitable for measurements of processing capacity.

Furthermore, three aspects of the followed procedure in this study could be improved in a following one. First, the students in this study completed the short-form CART as part of a lesson series, but passing or failing was not dependent on their performance on this test. We can therefore not be sure if maximum effort was applied for each item and by each student. In general, however, most students worked conscientiously. This could be due to the presence of the test taker, who was also a teacher, and to the fact that the students were explained why completing the test seriously was important. In addition, this test setting is representative for situations in which the test would be used formatively.

Second, this study deals with a Dutch translation of the short-form CART. Strictly speaking, therefore, we cannot be sure that the findings of the present study also apply to the original English version of the short-form CART. On theoretical grounds, however, there seems little reason to expect that translation should cause strong differences in the results: the triggered cognitive responses do not rely heavily on the language in which they are presented. Care was taken that information was ordered, explained and presented the same

as in the English version, and that complexity of wording and grammar was comparable. Furthermore, the translation was performed by a competent translator.

Third, the set-up of the test in Qualtrics (Qualtrics, Provo, UT) was such that it was possible to proceed to the next question without answering the previous one. This leads to missing data, which made imputation necessary in order to perform Mokken scale analysis. Imputation may to some degree compromise or bias the results of the analysis (Van der Ark et al., 2005), although the multiple imputation method which was applied here is the method which seems to cause the smallest discrepancies for Mokken scale analysis (Van Ginkel, Van der Ark, & Sijtsma, 2007). In addition, not forcing a response suits the underlying theory of processing underlying rationality: reaching a correct response requires not only ability, but also a willingness to engage. Furthermore, forced responses will very likely often be random guesses, especially in a test this extensive, and will as such not be manifestations of an underlying latent trait, which also would compromise the results. Although the percentage of missings increased as the test progressed, there is no evidence that scalability decreased accordingly: higher and lower *H* values (Table 2) are evenly spread out across the course of the test.

To conclude, two of the twelve subtests had sufficient discriminative power, and could therefore be used to validly order respondents on the assumed underlying trait. Three other subtests approached sufficient discriminative power. The remaining seven subtests could not be used for ordering respondents on the underlying trait. In its current form, then, not all short-form CART subtest sum scores could be meaningfully interpreted as a measure of a constituent part of rationality, which seems to limit its possibilities for use for assessment in higher education.

CRediT authorship contribution statement

Marion Tillema: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing. **Samantha Bouwmeester:** Conceptualization, Methodology, Formal analysis. **Peter Verkoeijen:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Anita Heijltjes:** Conceptualization, Resources, Writing - review & editing.

Acknowledgements

The short-form CART was obtained directly from the CART authors by means of a researcher request for use of the CART. We thank the CART authors for this opportunity.

The authors would like to thank the participating students and teachers of the Mechanical Engineering and Electrical and Computer Engineering at Avans University of Applied Sciences in 's-Hertogenbosch. A word of thanks also goes out to Geert Gooskens and Gabri Heinrichs for discussing previous versions of this work with the first author.

Appendix A. Examples of CART items translated into Dutch

Subtest A: Probabilistic and Statitistical Reasoning

Sample item original:

Consider the following hypothetical situation: A deck with 10 cards is randomly shuffled 10 separate times. The 10 cards are composed of 7 cards with the letter "A" on the down side and 3 cards with the letter "B" on the down side. Each time the 10 cards are reshuffled, your task is to predict the letter on the down side of the top card. Imagine that you will receive $100 for each downside letter you correctly predict, and that you want to earn as much money as possible. Indicate your predictions for each of the 10 shuffles:

I would predict ___ for Shuffle #1 A or B?
I would predict ___ for Shuffle #2 A or B?
I would predict ___ for Shuffle #3 A or B?
I would predict ___ for Shuffle #4 A or B?
I would predict ___ for Shuffle #5 A or B?
I would predict ___ for Shuffle #6 A or B?
I would predict ___ for Shuffle #7 A or B?
I would predict ___ for Shuffle #8 A or B?
I would predict ___ for Shuffle #9 A or B?
I would predict ___ for Shuffle #10 A or B?

Dutch translation:

Stel je voor: een pak van 10 kaarten wordt tien keer willekeurig geschud. Bij de 10 kaarten zitten 7 kaarten met de letter A op de onderkant en 3 kaarten met de letter B op de onderkant. Elke keer dat de 10 kaarten opnieuw worden geschud, heb jij de taak om de letter te voorspellen op de onderkant van de bovenste kaart. Stel je voor dat je 100 euro verdient voor elke juist voorspelde letter, en dat je zoveel mogelijk geld wilt verdienen. Geef aan welke voorspelling je bij elk van de schudbeurten doet:

Ik zou bij 1e keer schudden ___ voorspellen A of B?
Ik zou bij 2e keer schudden ___ voorspellen A of B?
Ik zou bij 3e keer schudden ___ voorspellen A of B?
Ik zou bij 4e keer schudden ___ voorspellen A of B?
Ik zou bij 5e keer schudden ___ voorspellen A of B?

Ik zou bij 6e keer schudden ___ voorspellen A of B?
Ik zou bij 7e keer schudden ___ voorspellen A of B?
Ik zou bij 8e keer schudden ___ voorspellen A of B?
Ik zou bij 9e keer schudden ___ voorspellen A of B?
Ik zou bij 10e keer schudden ___ voorspellen A of B?

Subtest B: Scientific Reasoning

Sample item original:

Each of the tickets below has a destination on one side and a mode of travel on the other side. Here is a rule: "**If 'Baltimore' is on one side of the ticket, then 'plane' is on the other side of the ticket.**" Your task is to decide which tickets you would need to turn over in order to find out whether the rule is being violated.

Destination: **Baltimore**		Destination: **Washington, D.C.**		Mode of Travel: **Train**		Mode of Travel: **Plane**	
Turn	Do not Turn	Turn	Do not Turn	Turn	Do not Turn	Turn	Do not Turn
O	O	O	O	O	O	O	O

Dutch translation:

Elk van de kaartjes hieronder heeft een bestemming op de ene zijde en een vervoersmiddel op de andere zijde. Hier is een regel: **"Als er 'Baltimore' op de ene kant staat, dan staat er 'vliegtuig' op de andere kant."** Het is jouw taak om te beslissen welke kaartjes moeten worden omgedraaid om erachter te komen of de regel wordt geschonden.

Bestemming: **Baltimore**		Bestemming: **Washington, D.C.**		Vervoermiddel: **Trein**		Vervoermiddel: **Vliegtuig**	
Omdraaien	Niet omdraaien	Omdraaien	Niet omdraaien	Omdraaien	Niet omdraaien	Omdraaien	Niet omdraaien
O	O	O	O	O	O	O	O

Subtest C: Reflection vs Intuition

Sample item original:
If it takes one minute to make each cut, how long will it take to cut a 25-foot wooden plank into 25 equal pieces?
Dutch translation:
Als het een minuut duurt om een snede te maken, hoe lang duurt het om een plank van 25 cm in 25 gelijke delen te zagen?

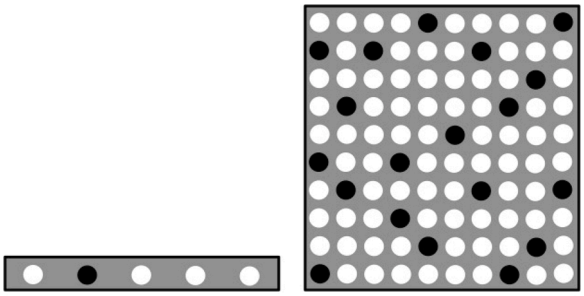
Subtest D: Syllogisms

Sample item original:
Premises:
Premise 1: All flowers are carbitops.
Premise 2: All tulips are carbitops.
Conclusion:
All tulips are flowers.
a. Conclusion necessarily follows from premises.
*b. Conclusion does not necessarily follow from premises.
Dutch translation:
Premissen:
Premisse 1: Alle bloemen zijn carbitopen.
Premisse 2: Alle tulpen zijn carbitopen.
Conclusie:
Alle tulpen zijn bloemen.
a. Conclusie volgt noodzakelijkerwijs uit de premissen.
*b. Conclusie volgt niet noodzakelijkerwijs uit de premissen.

Subtest E: Ratio Bias

Sample item original:

Assume that you are presented with two trays of black and white marbles (pictured below). The small tray contains **5** marbles. The large tray contains **100** marbles. The marbles inside each tray will be randomly mixed up, and you must draw out a single marble from one of the trays without looking. If you draw a black marble you win $5.

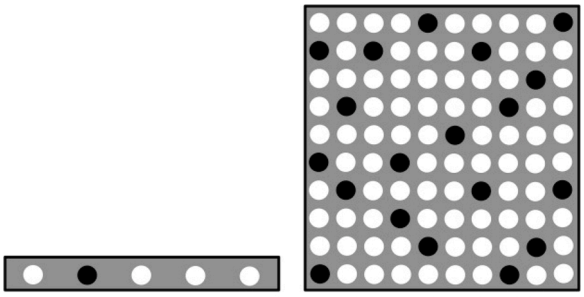


In a real situation, which tray would you prefer to select a marble from?

*a. Strongly prefer the small tray
*b. Moderately prefer the small tray
*c. Slightly prefer the small tray
d. Slightly prefer the large tray
e. Moderately prefer the large tray
f. Strongly prefer the large tray

Dutch translation:

Stel dat je twee bakken met zwarte en witte knikkers te zien krijgt (zie plaatje). De kleine bak bevat **5** knikkers. De grote bak bevat **100** knikkers. De knikkers in elke bak worden willekeurig gehusseld, en je moet uit één van de bakken bak één enkele knikker pakken zonder te kijken. Als je een zwarte knikker pakt, dan win je 5 euro.



Als dit een echte situatie was, uit welke bak zou je het liefst een knikker pakken?

*a. Sterke voorkeur voor de kleine bak.
*b. Enigszins een voorkeur voor de kleine bak.
*c. Beetje een voorkeur voor de kleine bak.
d. Beetje een voorkeur voor de grote bak.
e. Enigszins een voorkeur voor de grote bak.
f. Sterke voorkeur voor de grote bak.

Subtest F: Disjunctive Reasoning

Sample item original:

A food warehouse ships boxes of fresh and frozen strawberries. Assume that your job is to make sure these boxes are stacked properly. A box of fresh strawberries should not touch a box of frozen strawberries, because direct contact with the colder box will cause the fresh strawberries to spoil. You find a stack of three boxes of strawberries, where the top box contains fresh strawberries and the bottom box contains frozen strawberries. However, the middle box of strawberries has no label, and, thus contains either fresh or frozen strawberries. Is a box of fresh strawberries touching a box of frozen strawberries?

Fresh Strawberries
? Frozen Strawberries

*a. Yes
b. No
c. Cannot be Determined
Dutch translation:
Een supermarkt verscheept dozen met verse en bevroren aardbeien. Stel dat het jouw taak is om ervoor te zorgen dat de dozen goed worden gestapeld. Een doos verse aardbeien mag geen doos met bevroren aardbeien raken, want direct contact met een koudere doos zou de verse aardbeien doen bederven. Je vindt een stapel van drie dozen met aardbeien, waarvan de bovenste doos verse aardbeien bevat en de onderste doos bevroren aardbeien. De middelste doos heeft echter geen etiket en bevat dus ofwel verse, ofwel bevroren aardbeien. Is er in deze stapel een doos verse aardbeien die een doos met bevroren aardbeien raakt?

Verse aardbeien
? Bevroren aardbeien

*a. Ja
b. Nee
c. Kan niet worden vastgesteld

Subtest G: Probabilistic Numeracy

Sample item original:
Imagine that an unvaccinated person has a 10 % chance of getting the flu and that the flu vaccine is 80 % effective in preventing the flu. What are the chances that a person who has had the vaccine will still get the flu? ____
Dutch translation:
Stel: een niet-ingeënt person heeft een kans van 10 % om griep te krijgen en een inenting tegen griep is in 80 % van de gevallen effectief in het tegenhouden van griep. Wat is de kans dat een tegen griep ingeënt person toch griep krijgt? ____

Subtest H: Superstitious Thinking

Sample item original:
A person's thoughts can influence the movement of a physical object.
Dutch translation:
Het is voor mensen mogelijk om met hun gedachten het bewegen van een voorwerp te beïnvloeden.

Subtest I: Anti-Science Attitudes

Sample item original:
The fact that scientists often disagree about a topic shows that science involves the personal opinions of scientists more than actual evidence.
Dutch translation:
Het feit dat wetenschappers het vaak oneens zijn over een onderwerp laat zien dat wetenschap meer gaat over de persoonlijke opinies van wetenschappers dan over daadwerkelijk bewijs.

Subtest J: Conspiracy Beliefs

Sample item original:
High-level U.S. government operatives knew ahead of time that the 9/11 attack on the World Trade Center was about to occur.
Dutch translation:
Werknemers op het hoogste niveau van de overheid van de Verenigde Staten wisten van tevoren dat de aanval van 11 september 2001 op het World Trade Center ging gebeuren.

Subtest K: Dysfunctional Beliefs

Sample item original:

I worry a lot that I am unlikable.
Dutch translation:
Ik maak me vaak zorgen dat ik niet aardig ben.

Subtest L: Actively Open-Minded Thinking

Sample item original:
If a belief suits me then I am comfortable, it really doesn't matter if the belief is true. (R)
Dutch translation:
Als een overtuiging bij me past dan vind ik voel ik me prettig, het maakt dan niet zoveel uit of de overtuiging waar is. (R)

References

Bedford, A., Watson, R., Lyne, J., Tibbles, J., Davies, F., & Deary, I. J. (2010). Mokken scaling and principal components analyses of the CORE-OM in a large clinical sample. *Clinical Psychology & Psychotherapy, 17*(1), 51–62. https://doi.org/10.1002/cpp.649.

Bok, D. (2006). *Our underachieving colleges. A candid look at how much students learn and why they should be learning more*. Princeton, N.J: Princeton University Press.

Bonnefon, J. F. (2018). The pros and cons of identifying critical thinking with system 2 processing. *Topoi, 37*(1), 113–119. https://doi.org/10.1007/s11245-016-9375-2.

Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine, 84*(8), 1022–1028. https://doi.org/10.1097/ACM.0b013e3181ace703.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality & Safety, 22*(Suppl. 2), ii58–ii64. https://doi.org/10.1136/bmjqs-2012-001712.

Evans, J. S. B. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012.

Evans, J. S. B. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review, 31*(2-3), 86–102. https://doi.org/10.1016/j.dr.2011.07.007.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241. https://doi.org/10.1177/1745691612460685.

Fabio, R. A. (2005). Dynamic assessment of intelligence is a better reply to adaptive behavior and cognitive plasticity. *The Journal of General Psychology, 132*(1), 41–66.

Facione, P. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (the Delphi Report)*.

Facione, P. A., & Facione, N. C. (1992). *The California critical thinking skills test*. California: Academic Press.

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *The Quarterly Journal of Experimental Psychology, 71*(5), 1188–1208. https://doi.org/10.1080/17470218.2017.1313283.

Halpern, D. F. (1999). Teaching for critical thinking: Helping college students develop the skills and dispositions of a critical thinker. *New Directions for Teaching and Learning, 80*, 69–74.

Halpern, D. F. (2010). Halpern critical thinking assessment. *SCHUHFRIED (Vienna test system): Moedling, Austria* (p. 1). Http://www.Schuhfried.com/vienna-Test-System-vts/all-Tests-fromaz/test/hcta-Halpern-Critical-Thinking-Assessment.

Janssen, E. M., Verkoeijen, P. P., Heijltjes, A. E., Mainhard, T., van Peppen, L. M., & van Gog, T. (2020). Psychometric properties of the actively open-minded thinking scale. *Thinking Skills and Creativity*. , Article 100659. https://doi.org/10.1016/j.tsc.2020.100659.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Klaczynski, P. A., & Felmban, W. (2020). Effects of thinking dispositions, general ability, numeracy, and instructional set on judgments and decision-making. *Psychological Reports, 123*(2), 341–370. https://doi.org/10.1177/0033294118806473.

Koehler, D. J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. *Heuristics and biases: The psychology of intuitive judgment* (pp. 686–715).

Leach, S., Immekus, J. C., French, B. F., & Hand, B. (2020). The factorial validity of the cornell critical thinking tests: A multi-analytic approach. *Thinking Skills and Creativity, 37*. https://doi.org/10.1016/j.tsc.2020.100676.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. De Gruyter Mouton.

OECD. (2018). *Teaching, assessing and learning creative and critical thinking skills in education*. http://www.oecd.org/education/ceri/assessingprogressionincreativeandcriticalthinkingskillsineducation.htm.

Parker, A. M., & Fischhoff, B. (2005). Decision-making competence: External validation through an individual-differences approach. *Journal of Behavioral Decision Making, 18*(1), 1–27. https://doi.org/10.1002/bdm.481.

Qualtrics, P., UT, USA. https://www.qualtrics.com.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org.

Rachlinski, J. J. (2004). Heuristics, biases, and governance. *Blackwell handbook of judgment and decision making* (p. 567).

Rear, D. (2019). One size fits all? The limitations of standardised assessment in critical thinking. *Assessment and Evaluation in Higher Education, 44*(5), 664–675. https://doi.org/10.1080/02602938.2018.1526255.

RStudio Team. (2020). *RStudio: Integrated development for R*. Boston, MA: R Studio, PBC. http://www.rstudio.com/.

Siegel, H. (1989). The rationality of science, critical thinking, and science education. *Synthese, 80*(1), 9–41. http://www.jstor.org/stable/20116665?origin=JSTOR-pdf.

Sijtsma, K. (2009). Over misverstanden rond Cronbachs alfa en de wenselijkheid van alternatieven. [On misunderstandings surrounding Cronbach's alpha and the desirability of alternatives]. *De Psycholoog, 44*(11), 561. https://research.tilburguniversity.edu/en/publications/388365fc-5dc1-4ff2-a351-92f63825e746.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). SAGE.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2017). *The rationality quotient*. Cambridge, MA: MIT Press.

Tversky, A., & Kahneman, D. (1974). *Judgment under uncertainty*. Heuristics and biases. *Science, 185*(4157), 1124–1131.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1–67.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of statistical software, 20(11), 20*(11), 1–19.

Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (Eds.). (2005). *New developments in categorical data analysis for the social and behavioral sciences*. Psychology Press.

Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research, 42*(2), 387–414.

Van Peppen, L. M., Verkoeijen, P. P., Heijltjes, A. E., Janssen, E. M., Koopmans, D., & Van Gog, T. (2018). Effects of self-explaining on learning and transfer of critical thinking skills. November. In *Frontiers in education* (Vol. 3, p. 100). https://doi.org/10.3389/feduc.2018.00100. Frontiers.

Wason, P. C., & Evans, J. S. B. (1974). Dual processes in reasoning? *Cognition, 3*(2), 141–154.