CHAPTER 2

Development and validity of a Dutch version of the Remote Associate Task:

An item-response theory approach

Akbari Chermahini, S., Hickendorff, M., & Hommel, B. (submitted). Development and validity of a Dutch version of the Remote Associate Task: An Item Response Theory approach.

ABSTRACT

The Remote Associates Test (RAT) developed by Mednick (1967) is known as a valid measure of creative convergent thinking.We developed a 30-item version of the RAT in Dutch language with high internal consistency (Cronbach's alpha =0.85) and applied both Classical Test Theory and Item ResponseTheory (IRT) to provide measures of item difficulty and discriminability, construct validity, and reliability. IRT was further used to construct a shorter version of the RAT, which comprises of 22 items but still shows good reliability and validity—as revealed by its relation to Raven's Advanced Progressive Matrices test, another insight-problem test, and Guilford's Alternative Uses Test.

INTRODUCTION

Most researchers agree that creativity is the ability to generate behavior and behavioral outcomes that are unique, useful, and productive (Sternberg, et al, 1996). Therefore, creativity is considered as a performance or ability, manifested in original, valuable, and socially accepted ideas, products, or works of art. The creativity level of an individual can be assessed by means of performance measures derived from creative thinking tasks. Guilford (1967), who can be considered the founder of modern creativity research, drew a distinction between convergent and divergent thinking. Convergent thinking aims for a single, highly constrained solution to a problem, whereas divergent thinking involves the generation of multiple answers to an often loosely defined problem.

Influenced by Guilford's suggestions to distinguish convergent and divergent thinking, many creativity measures have been developed, such as Guilford's Alternative Uses Test, considered to assess divergent thinking, and Mednick's Remote Associates Test (RAT; Mednick, Mednick, & Mednick, 1964), considered to assess convergent thinking. The latter was designed in accordance with S. Mednick's (1962) associative theory of creativity. According to this theory, the creative thinking process consists inusing associative elements to create new combinations which either meet specified requirements or are in some way useful.

The test aimed at measuring creative thought without requiring knowledge specific to any particular field. Two college-level versions of the test were developed, each consisting of 30 items (Mednick, 1968; Mednick & Mednick, 1967). Each item consists of three words that can be associated in a number of ways, such as by forming a compound word or a semantic association. "Creative thought" is required to find a correct solution because the first and most obvious solution is often not correct, so that more remote connections need to be retrieved in order to relate the three words to each other. Even though this arguably introduced an aspect of divergent thinking, the basic structure of the RAT (finding a highly constrained, single solution) fits rather well with Guilford's (1967) concept of convergent thinking. Notwithstanding Guilford's distinction, in most studies of problem solving and creative thinking the RAT has been used as a test of general creativity (e.g., Ansburg, 2000; Beeman & Bowden, 2000; Bowers, Regehr, Balthazard, & Parker, 1990; Dallob &

Dominowski, 1993; Dorfman, Shames, & Kihlstrom, 1996; Schooler & Melcher, 1995; Shames, 1994; Smith & Blankenship, 1991). The RAT has also been employed in a wide range of research including studying psychopathologies (e.g., Fodor, 1999), success and failure experiences (e.g., Vohs & Heatherton, 2001), affect (e.g., Mikulincer & Sheffi, 2000).

Performance on the RAT is known to correlate with performance on classic insight problems (e.g., Dallob & Dominowski, 1993; Schooler & Melcher, 1995; Öllinger et al. 2008; Ansbug, 2000; Daialey, 1978), suggesting that at least some items in the RAT reflect insight. The materials used in the test involve verbal associative habits that could reasonably be assumed to be familiar to almost all individuals brought up in the United States, especially in the English speaking part of the US culture. However, it has been noted that the RAT is rather difficult for non-native speakers of English (e.g., Estrada, Isen& Young, 1994). Several non-English versions have therefore been developed: Hebrew, Japanese, and Jamaican (Baba, 1982; Hamilton, 1982; Levin & Nevo, 1978), but to our knowledge there is no Dutch version of this test available. Therefore, the aim of the current study was to develop a Dutch version of the RAT: a short, reliable, and valid measurement instrument to measure convergent thinking in the Dutch language. To do so we first developed and administered 30 Dutch RAT-like items. Next, we used Item Response Theory (IRT) to evaluate the psychometric properties of this 30-item test, and to shorten the test with the least possible loss of psychometric quality and information. To validate this short version, we related the RAT measures to measures from two other tasks that are assumed to assess aspects of convergent thinking: the Raven's Advanced Progressive Matrices test (Raven, 1965), which is also considered to provide an estimate of fluid intelligence, and an insight-problem test. Finally, we contrasted RAT measures with estimates of divergence-thinking performance derived from Guilford's Alternative Uses Test.

METHOD

Participants and Procedure

Participants were students from Leiden University, the Netherlands. All of them were

native speakers of Dutch. The sample consisted of 158 participants (133 females and 25 males). Their age ranged from 18 to 32, with a mean of 20.4 (SD=2.9). They were tested individually in 60-min sessions, in which they worked through three paper-and-pencil-type tests (the Dutch RAT, an insight problem test, and the Alternative Uses Task, all described below), and a computer version test of Raven's Advanced Progressive Matrices.

Instrument

Remote Associate Test (RAT)

Of the original, English RAT (Mednick, 1962) two college–level versions have been constructed, each consisting of 30 items. For each item, three words are presented and the participant is required to identify the (fourth) word that connects these three seemingly unrelated words (e.g., "bass, complex, sleep", where the solution is "deep"). The solution word for each item can be associated with the words of the triad in various ways, such as synonymy, formation of a compound word, or semantic association. The link between the words is associative and does not follow common rules of logic, concept formation, or problem solving. Hence, with all items of the test the solution word is a remote, uncommon associate of each of the stimulus words, requiring the respondent to work outside of these common analytical constraints. The score is determined by the number of correct answers given in a particular time.

We constructed a Dutch version of the RAT as follows: First, native Dutch-speaking staff members of the psychological department of Leiden University were consulted to construct 50 sets of words. Each set consisted of three words that were associated with a solution word. Next, a group of students from Leiden University (all native Dutch speakers) were asked to respond to these 50 items, providing a check for strange or saliently uncommon items. Based on this screening process, 30 items were chosen. Finally, a separate group of 158 students—the actual participants of this study—were asked to respond to the 30 item within 10 minutes.

Raven's Advanced Progressive Matrices

Raven's Advanced Progressive Matrices (APM: Raven, 1965) test is considered to assess insight and has been constructed to provide a language-independent estimate of fluid

intelligence and Spearman's g. We used 36 items on which participants worked for 25 minutes. Each item of this test consists of a visual pattern with one piece missing, which participants are to identify from a set of alternatives. The items get progressively harder and are assumed to need increasingly more cognitive capacity.

Insight Problem

An insight problem is a problem that requires participants to shift their perspective and view the problem in a novel way to achieve the solution. According to the domain-specific theory (see Baer in Runco, 1999), insight problems can be divided into coherent subcategories such as verbal, mathematical, and spatial insight problems (Dow & Mayer 2004). The insight problem test in this study (see Appendix) consisted of three questions that included all three subcategories of insight problems: a verbal and a spatial problem (both adopted from Metcalfe, 1986), and a mathematical problem (adopted from Sternberg & Davidson, 1982). Participants were asked to do the test in 15 minutes. The total number of correct responses was used as score.

Alternative Uses Task

In this task (based on Guilford, 1967), participants were asked to list as many possible uses for three common household items (*brick*, *shoe*, and *newspaper*) as they can within 10 minutes. Scoring comprised of four components:

Originality: Each response is compared to the total amount of responses from all of the participants. Responses that were given by only 5% of the group counted as unusual (1 point) and responses given by only 1% of them count as unique (2 points).

Fluency: The total of all responses.

Flexibility: The number of different categories used.

Elaboration: The amount of detail; e.g., "a doorstop" counts 0, whereas "a door stop to prevent a door slamming shut in a strong wind" counts 2 (1 point for explanation of door slamming and another for further detail about the wind).

Data analysis

Psychometric theory offers two approaches to evaluate the design, analysis, and scoring of tests: Classical Test Theory (CTT) and Item Response Theory (IRT; see Embretson & Reise, 2000). Both theories allow predicting outcomes of psychological tests by identifying parameters of item difficulty and the ability of test takers, and both provide measures to assess the reliability and validity of psychological tests.

CTT is widely used as a method of analysis in evaluating tests but it has some limitations. First, the observed total score is item dependent. That is, if two participants complete different tests that measure the same construct, the meaning of their total scores depend on the difficulty of the items in their respective tests. Often observed side-effects are floor and ceiling effects. Second, item statistics or the difficulty level and item discrimination are examinee dependent. That is, the commonly used CTT-statistic for difficulty level, the *P*-value (probability correct), depends on the ability level of the sample of test takers: the *P*-value will be higher in samples with high than with low ability levels. Moreover, the CTT-statistic for the discrimination of an item, the item-rest-correlation, will be highest if participants have around 50% chance to answer the item correctly. So, these statistics also depend on the specific sample of test takers.

IRT overcomes these limitations of CTT. In IRT, each item in a test has its own characteristic curve which describes the probability of answering the item correctly depending on the test taker's ability (Kaplan & Saccuzzo, 1997). One of the advantages of using IRT over CTT is IRTs sample-independent nature of its results. This means that item parameters are invariant when computed from different groups of different ability levels. As a result, the same measurement scale can be used in different groups of participants, and groups as well as individuals can be tested with a different set of items, appropriate to their ability levels. Their scores will be directly comparable (Anastasi & Urbina, 1997). Because of these advantages, we applied IRT modeling in this study in evaluating item and test properties to judge the test's reliability and validity. IRT asserts that the easier the question, the more likely a participant will be able to respond to it correctly, and the more able the participant, the more likely he or she will be able to answer the question correctly as compared to a student who is less able. In IRT models, it is assumed that there exists a latent (unobserved) ability scale, usually called θ, that underlies performance on a set of items. The

probability that a person answers an item correctly is modeled as function of this person's latent ability, and a set of item parameters. The probability of a correct answer on an item increases with higher latent ability, following an S-shaped curve bounded by 0 and 1: the *Item Characteristic Curve*. There are three common item parameters: the difficulty, discrimination, and guessing parameter. The *difficulty* or location parameter manages the curve's point of inflection (the level of θ yielding a 50% probability of a correct answer), the *discrimination* parameter determines its slope, and the *guessing* parameter represents the lower asymptote.

Item characteristic curves provide important and useful information about item properties. IRT can also be used to study item and test *information functions*. *Item Information Curves* (or functions) indicate the range over θ where an item is best at discriminating among individuals. More information, determined by the item's discrimination parameter, indicates higher accuracy or reliability for measuring a person's trait level. Item information can be used to select a set of items that together provide much information on a desired range of latent the ability scale. *The Test Information Curve* (or function) indicates the amount of information (i.e., reliability) provided by the scale over the range of the construct continuum. The test information curve is simply the sum of the item information curves of the items in the test. *The Standard Error of Measurement* is reciprocally related to the test information function, and evaluates the accuracy of the test to measure people at different levels along the ability continuum.

RESULTS

Classical Test Theory

The mean RAT total score was 8.94 (SD =5.21). Internal consistency of the scale was determined using Cronbach's alpha as a function of the mean inter-item correlations among the 30 dichotomously scored items. The high alpha value (0.85) of the scale is a sign of very good internal consistency with this sample, indicating that the items are consistent in measuring the underlying construct. The first two columns in Table 1 show, for each item, the total probability correct in the sample (ranging from .02 to .72) and the item-rest correlations (ranging from .09 to .65). In general, the 30 items appear rather difficult, and all

items are positively related to the overall test score, although this relation is stronger for some items than for others.

Item Response Theory

Two IRT models were compared in the analyses. A one-parameter logistic (1PL) model was specified in which item difficulties were freely estimated but item discriminations were constrained to be equal and item lower asymptotes (guessing parameter) were fixed at 0. A two-parameter logistic (2PL) model was specified in which item difficulties and discriminations were freely estimated but again lower asymptotes were fixed at 0. Because of the open-ended nature of the Remote Association Task items, it makes no sense to apply the guessing parameter, so the three- parameter model (3PL), which freely estimates difficulties, discriminations, and lower asymptotes is not useful here. The two IRT models (1PL and 2PL) were fit with Rizopoulos's (2006) IRT program for R language (R Development Core Team, 2009) (In this program, it is assumed that θ follows a normal distribution with mean zero and standard deviation 1). Model fit statistics are presented in Table 2.

Likelihood ratio tests revealed that the 2PL model provided significantly better fit than the 1PL model, LRT (29) = 68.21, $p<0.001$. The AIC-values (lower values imply better trade-off between statistical model fit and model complexity) also point to the 2PL model as the best fitting one. Item parameter estimates and item fit statistics for the 2PL model are presented in the last four columns of Table 1, with items ordered with respect to increasing difficulty level. The resulting Item Characteristic Curves are depicted in Figure 1.

Table 1: Classical Test Theory (CTT) Statistics, and Item response Theory (IRT) Item Parameter Estimates (With Standard Errors) and Fit Statistics for the Two-Parameter Logistic (2PL) Model of 30-Item RAT.

Item		CTT-Statistics		IRT-Item parameters		IRT-Item fit	
		Probability correct	*Item-Rest Correlation*	Difficulty	Discrimination	χ^2	*Boot strapped p-value*
1	bar/jurk/glas	0.72	0.65	-0.58 (0.12)	4.08 (1.13)	4.82	0.78
2	room/vloot/koek	0.59	0.31	-0.46 (0.24)	0.87 (0.22)	21.1	0.01
3	kaas/land/huis	0.63	0.51	-0.45 (0.17)	1.53 (0.32)	5.75	0.74
4	vlokken/ketting/pet	0.60	0.48	-0.34 (0.16)	1.59 (0.32)	3.83	0.97
5	val/melon/lelie	0.58	0.51	-0.25 (0.15)	1.69 (0.35)	10.4	0.31
6	vis/mijn/geel	0.56	0.48	-0.19 (0.16)	1.44 (0.30)	4.66	0.85
7	achter/kruk/mat	0.51	0.42	-0.03 (0.17)	1.25 (0.28)	13.63	0.12
8	worm/kast/legger	0.48	0.46	0.10 (0.15)	1.48 (0.32)	4.31	0.94
9	water/schoorsteen/lucht	0.46	0.52	0.16 (0.13)	1.93 (0.41)	12.75	0.18
10	trammel/beleg/mes	0.37	0.46	0.49 (0.14)	1.72 (0.38)	9.86	0.18
11	hond/druk/band	0.38	0.46	0.50 (0.17)	1.37 (0.32)	12.01	0.15
12	goot/kool/bak	0.35	0.46	0.58 (0.16)	1.58 (0.36)	7.92	0.52
13	controle/plaats/gewicht	0.36	0.45	0.58 (0.18)	1.33 (0.31)	9.61	0.36
14	kolen/land/schacht	0.32	0.51	0.60 (0.13)	2.44 (0.61)	4.55	0.84
15	schommel/klap/rol	0.37	0.33	0.63 (0.21)	1.07 (0.27)	10.03	0.30
16	kamer/masker/explosie	0.26	0.35	1.12 (0.28)	1.16 (0.32)	9.37	0.27
17	nacht/vet/licht	0.17	0.36	1.46 (0.31)	1.41 (0.40)	15.11	0.06
18	arm/veld/stil	0.20	0.24	2.04 (0.68)	0.74 (0.26)	10.6	0.27
19	olie/pak/meester	0.22	0.23	2.23 (0.83)	0.62 (0.24)	8.24	0.46
20	school/ontbijt/spel	0.04	0.29	2.45 (0.61)	1.80 (0.68)	11.9	0.14
21	kop/boon/pause	0.11	0.22	2.49 (0.79)	0.94 (0.34)	13.64	0.12
22	licht/dromen/maan	0.15	0.22	2.49 (0.84)	0.79 (0.30)	6.95	0.57
23	deur/werk/kamer	0.05	0.24	2.81 (0.83)	1.26 (0.49)	5.14	0.65
24	ga/daar/dag	0.11	0.22	2.98 (1.09)	0.78 (0.32)	13.08	0.13
25	strijkijzer/schip/trein	0.02	0.20	3.24 (0.99)	1.54 (0.67)	6.7	0.38
26	man/lijm/ster	0.12	0.21	3.30 (1.39)	0.64 (0.30)	9.92	0.21
27	bed/zee/school	0.02	0.21	3.42 (1.12)	1.42 (0.64)	17.72	0.05
28	riet/klontje/hart	0.10	0.18	3.43 (1.43)	0.69 (0.32)	2.84	0.98
29	palm/familie/huis	0.04	0.16	3.70 (1.44)	0.98 (0.46)	4.01	0.80
30	grond/vis/geld	0.08	0.09	5.29 (3.38)	0.49 (0.33)	8.25	0.47

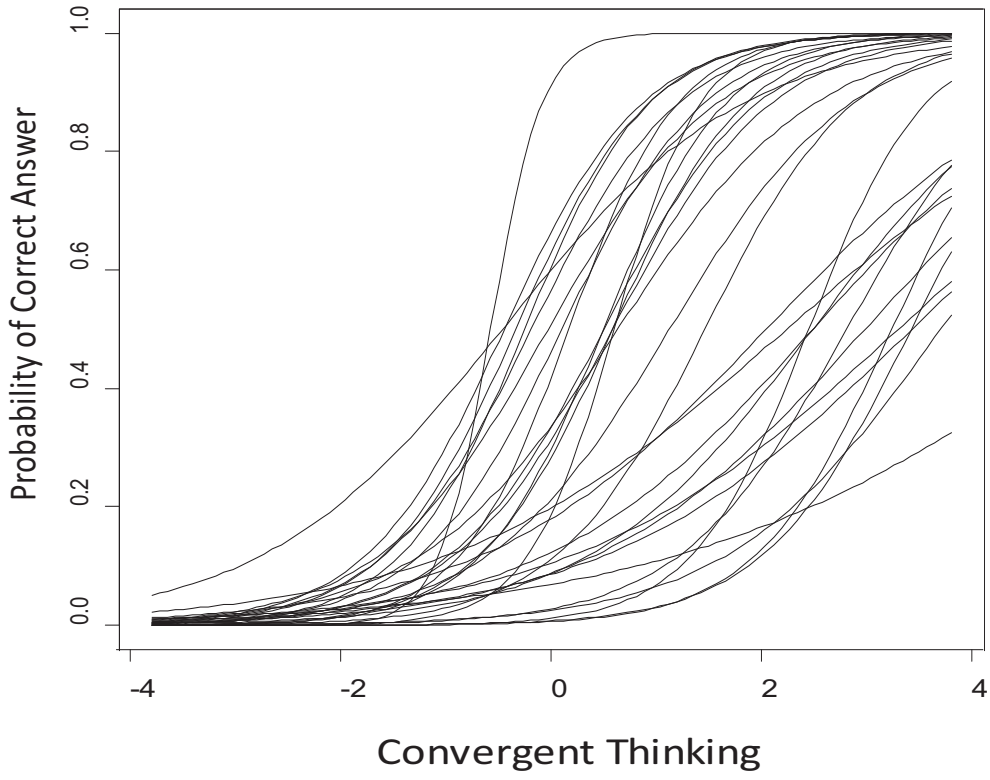

Figure 1: Item Characteristic curves for all 30 items of Remote Association Task. Functions were produced with a 2PL (two-parameter logistic)Item Response Theory model.

Table 2: Fit Statistics for the 1PL and 2PL Logistic Models of 30-item test

Test	Model	In*L*	No. of parameters	AIC	BIC
30- item	1PL	- 069.32	31*	4200.65	4295.59
	2PL	- 035.22	60	4190.43	4374.19

Note. 1PL = one-parameter logistic model; 2PL = two-parameter logistic model; In*L* = log-likelihood;

AIC = Akaike information coefficient

BIC = Bayesian information coefficient

*Thirty item difficulty parameters plus a common discrimination parameter.

Table 1 shows that the difficulty levels range between -.58 (fairly easy item) and 5.29 (extremely hard item). Only 7 items have a difficulty level that is below 0 (an item with difficulty parameter 0 would be solved correctly with 50% probability by a participant with average ability level); while 23 items have a difficulty level higher than 0. In particular, 13 items are very difficult with a difficulty level above 2.00, meaning that only participants with $\theta >$ 2.00 have a probability of 50% or higher to answer these items correctly. Because it is rather unlikely that there are many individuals with such high ability levels (based on the standard normal distribution, only 2.5% of the participants have a θ-level of at least 1.96), it is not necessary that there are so many difficult items in this test. Therefore, 7 of these items, having a low discrimination parameter, were selected as candidates for removal. Moreover, one item (item 2) showed significant misfit to the 2PL model ($p<.01$), and was therefore also removed from the test.

Thus, 22 items were selected as the best items in terms of difficulty and discrimination levels. Another set of 1PL and 2PL models were carried out to analyze the data of the 22 selected items. Model fit statistics are presented in Table 3. Likelihood ratio tests revealed that also for the 22 selected items the 2PL model provided significantly better fit than did 1PL model, LRT (21) = 40.97, $p<0.01$.

Table 3: Fit Statistics for the 1PL and 2PL Logistic Models of 22-item test

Test	Model	In*L*	No. of parameters	AIC	BIC
22-item	1PL	- 626.85	23 *	3299.71	3370.15
	2PL	- 606.37	44	3300.73	3435.49

* Twenty-two item difficulty parameters plus a common discrimination parameter.

Item parameter estimates and fit statistic for the 2PL model are presented in Table 4 and Figure 2.Although there is still an overrepresentation of the more difficult items on this 22-item scale, the imbalance is much less extreme. In addition, the test was shortened by 27% of its length without losing much psychometric information, as comes forward from the test information curves of the 30-item test (Figure 3a) and the 22-item test (Figure 3b). More specifically, in the θ-range that comprises of approximately 95% of the participants (between -2 and +2) the test information decreased by only 10% by dropping 8 of the 30 items. Finally, the item fit statistics (Table 4) show that there are no items that show significant misfit to the 2PL model anymore. In conclusion, compared to the 30-item test, the 22-item test shows only minor loss in information, but a substantial shortening of the test. Cronbach's alpha of the 22-item test is still high at 0.84.

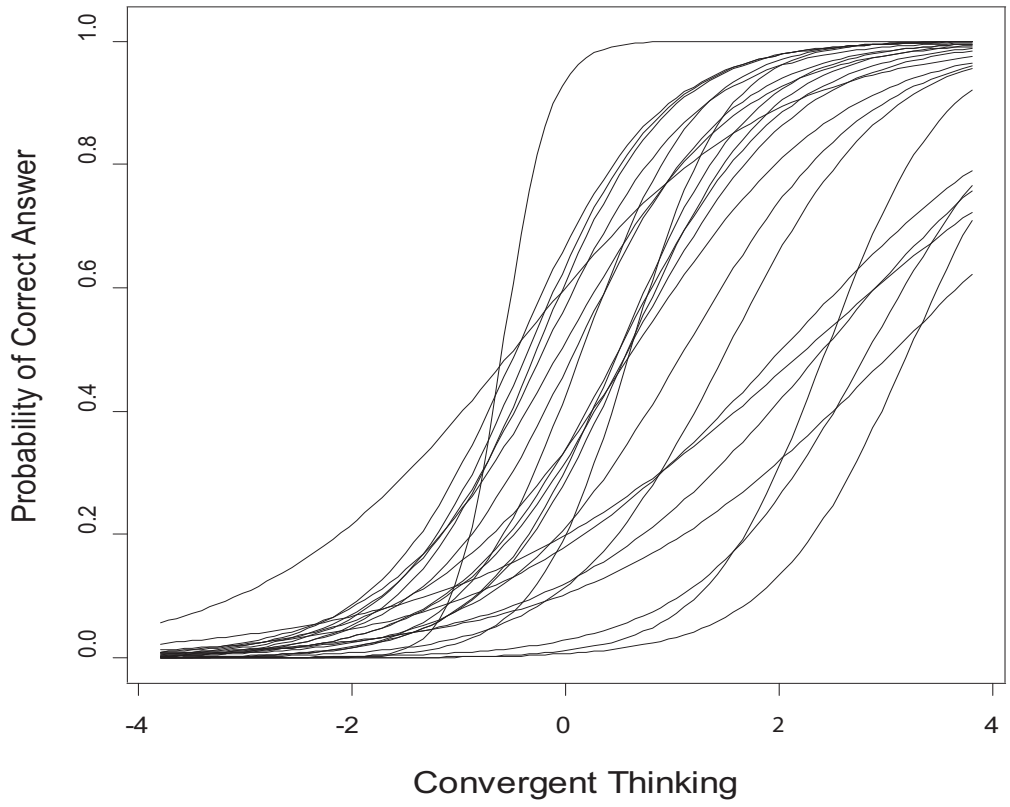

Figure 2: Item Characteristic curves for all 22 items of Remote Association Task. Functions were produced with a 2PL(two-parameter logistic) Item Response Theory model.

Table 4: Item response Theory (IRT) Item Parameter Estimates (With Standard Errors) and Fit Statistics for the Two-Parameter Logistic (2PL) Model of 22-Item RAT.

Item		IRT- Item parameters		IRT- Item fit	
		Difficulty	Discrimination	χ^2	*Bootstrapped p-value*
1	Bar/jurk/glas	-0.60 (0.12)	4.15 (1.25)	5.77	0.59
2	Kaas/land/huis	-0.45 (0.16)	1.61 (0.34)	7.64	0.56
3	Vlokken/ketting/pet	-0.35 (0.15)	1.59 (0.33)	6.54	0.71
4	Val/melon/lelie	-0.27 (0.15)	1.69 (0.35)	10.27	0.17
5	Vis/mijn/geel	-0.20 (0.16)	1.45 (0.31)	2.83	0.99
6	Achter/kruk/mat	-0.04 (0.17)	1.24 (0.28)	8.77	0.43
7	Worm/kast/legger	0.09 (0.15)	1.43 (0.31)	2.32	1.00
8	Water/schoorsteen/lucht	0.15 (0.13)	1.88 (0.39)	9.8	0.25
9	Trammel/beleg/mes	0.48 (0.15)	1.72 (0.38)	8.27	0.38
10	Hond/druk/band	0.49 (0.17)	1.34 (0.31)	7.55	0.57
11	Controle/plaats/gewicht	0.59 (0.18)	1.29 (0.31)	5.98	0.72
12	Goot/kool/bak	0.59 (0.17)	1.48 (0.34)	8.7	0.45
13	Kolen/land/schacht	0.61 (0.14)	2.20 (0.53)	9.3	0.31
14	Schommel/klap/rol	0.62 (0.21)	1.09 (0.27)	12.25	0.22
15	Kamer/masker/explosie	1.12 (0.28)	1.15 (0.31)	7.05	0.60
16	Nacht/vet/licht	1.59 (0.34)	1.31 (0.37)	8.48	0.45
17	Arm/veld/stil	2.02 (0.64)	0.75 (0.26)	5.5	0.74
18	Olie/pak/meester	2.28 (0.86)	0.61 (0.24)	5.21	0.84
19	School/ontbijt/spel	2.60 (0.66)	1.64 (0.61)	6.9	0.44
20	Deur/werk/kamer	2.86 (0.85)	1.23 (0.47)	4.86	0.83
21	Strijkijzer/schip/trein	3.28 (1.02)	1.51 (0.68)	7.37	0.44
22	Man/lijm/ster	3.49 (1.19)	1.38 (0.64)	18.21	0.11

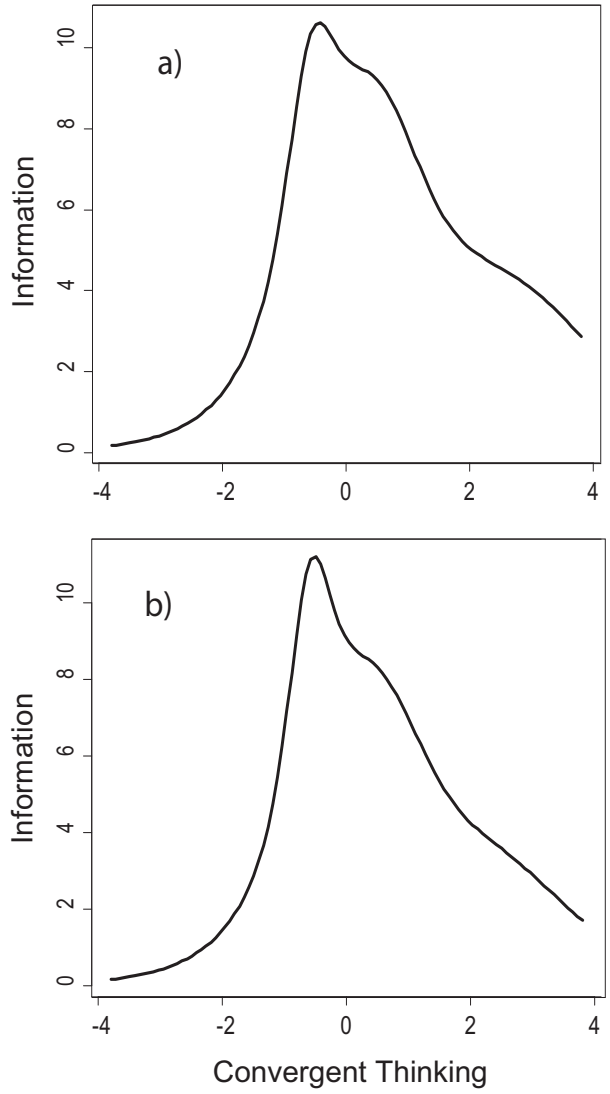

Figure 3: Test information function plotted against convergent thinking as a normally distributed latent factor for 30-item (a), and 22-item (b) tests.

Convergent and Discriminant Validity

Convergent validity has been defined as "how well the construct's measurement positively correlates with different measurements of the same construct" (Hair, 2003). Discriminant validity is the degree to which scores on a test do not correlate with scores from other tests that are not designed to measure the same construct.

In IRT, subjects answering the same number of items correctly typically do not have the same ability estimates unless they have answered exactly the same set of items correctly. Therefore, in this part of the research, individual scores on the RAT were derived from the 22-item IRT scale model parameters. We used Expected a Posteriori (EAP; e.g., Embretson, & Reise, 2000) scoring to obtain an ability estimate for each participant.

Convergent validity was evaluated using correlations between the scores derived from RAT (22-item), Raven's Advanced Progressive Matrices, and the Insight Problems—which were all assumed to represent aspects of convergent-thinking performance. To examine discriminant validity, correlations between RAT scores and the four scales of the Alternative Uses Task (a test to assess divergent thinking) were calculated.

As Table 5 shows, the correlations between RAT scores and both Raven scores and Insight Problem scores are significant. As both the Raven and the Insight problem tasks are assumed to assess aspects of convergent thinking—which explains why they also correlate with each other, this provides evidence for a substantial convergent validity of the developed RAT. Moreover, the results in Table 5 show that the RAT score correlate with none of the four AUT scores, which is consistent with Guilford's (1967) distinction between convergent and divergent thinking and demonstrates the discriminative validity of our version of the RAT.

Table 5: Coefficients and significance levels (** for p<.01 and * for p<.05) for tests of correlation between Remote Association Task (RAT: 22-item), Insight Problems (IP), Raven's Advanced Progressive Matrices (Raven), and Alternative Uses Task (AUT, FLU=fluency, FLE=flexibility, ORI=originality, ELA=elaboration).

	RAVEN	IP	AUT- FLU	AUT-FLE	AUT-ORI	AUT-ELA
RAT (22-item)	**0.47****	**0.39****	-0.07	0.07	-0.01	-0.13
RAVEN		**0.32****	-0.14	-0.05	-0.05	-0.08
IP			-0.12	0.02	0.02	-0.08

DISCUSSION

The aim of this study was to develop a short, reliable, and valid Dutch version of Mednick's (1967) RAT, which is widely used and considered a reliable measure of creative (convergent) thinking. To do so, we collected and analyzed data from a sample of Dutch university students. The CTT analysis revealed that the original 30-item test has high internal consistency (Cronbach's alpha =.85). The IRT analysis allowed us to reduce the 30-item set to a more efficient 22-item version that proved to be a high-quality instrument. The items were most consistent with a 2PL RIT model and they had unique discrimination and difficulty parameters. As expected, the Dutch 22-item RAT score was related to fluid intelligence scores, as measured by the Raven, and insight problem solving, as assessed by our 3-domain compound task, but not to divergent thinking. These findings provide strong evidence for the convergent and discriminant validity of our task version, respectively, which result in good construct validity. Furthermore, these findings encourage the use of the test as a good measure of creative convergent thinking.

Although the present study provides encouraging results, our sample (n=158) was not very large and restricted to university students. This is likely to be sufficient for standard experimentation, which usually considers student at participants, but may not provide a solid basis for investigating a more diverse population including children and elderly participants, or participants with a more diverse educational background. Accordingly, we regard the present evidence for the validity of the test preliminary. Although the 30-item is reliable and has high internal consistency, we recommend the 22-item version for most studies, as it is less time-consuming and does not contain very difficult and low-discriminant items. However, it is possible that studies in highly gifted individuals benefit from the inclusion of the highly difficult items that we excluded in the present study.

IRT-based models have been studied extensively and widely implemented in educational measurement for investigating the properties of tests, items, and examinees. IRT analyses can contribute to the improvement of the assessment instruments, ultimately enhancing the validity of the instrument. As far as we know, our study is the first to apply IRT to validate the RAT. To summarize, the Dutch 22-item version of the RAT developed in the present study provides a convenient and rather efficient test to measure convergent thinking with an instrument that possesses satisfactory psychometric properties.

REFERENCES

Anastasi, A., & Urbina, S. (1982). Psychological testing. *New York*.

Ansburg, P. I. (2000). Individual differences in problem solving via insight. *Current Psychology, 19*(2), 143-146.

Baer, J. (1999). Domains of creativity, In M. A. Runco, & S. R. Pritzker,(Ed.),*Encyclopedia of Creativity, 1,(pp.* 591 – 596). San Diego, CA: Academic Press.

Baker, F. B. (2004). *Item response theory: Parameter estimation techniques* (Vol. 176): CRC.

Beeman, M. J., & Bowden, E. M. (2000). The right hemisphere maintains solution-related activation for yet-to-be-solved problems. *Memory & Cognition, 28*(7), 1231-1241.

Bowden, E. M., & Jung-Beeman, M. (2003). Aha! Insight experience correlates with solution activation in the right hemisphere. *Psychonomic Bulletin & Review, 10*(3), 730.

Bowers, K. S., Regehr, G., Balthazard, C., & Parker, K. (1990). Intuition in the context of discovery. *Cognitive psychology, 22*(1), 72-110.

Dow, G.T. & Mayer, R.E. (2004). Teaching students to solve insight problems. Evidence for domain specificity in training. *Creativity Research Journal, 16*(4),389-402

Dallob, P., & Dominowski, R. (1993l). *Erroneous solutions to verbal insight problems: Effects of highlighting critical material.* Paper presented at the Meeting of the Western Psychological Association

Dorfman, J., Shames, V. A., & Kihlstrom, J. F. (1996). Intuition, incubation, and insight: Implicit cognition in problem solving. In D. M. Underwood Geoffrey (Ed.), *Implicit cognition* (pp. 257-296). Oxford: The Oxford University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*: Lawrence Erlbaum.

Fodor, E. M. (1999). Subclinical inclination toward manic-depression and creative performance on the Remote Associates Test. *Personality and individual differences, 27*(6), 1273-1283.

Kaplan, R. M., & Saccuzzo, D. P. (2008). *Psychological testing: Principles, applications,*

and issues. Pacific Grove Wadsworth Pub Co.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*: L. Erlbaum Associates Hillsdale, NJ.

Mednick, S. (1962). The associative basis of the creative process. *Psychological review, 69*(3), 220-232.

Mednick, S. A. (1968). The Remote Associates Test. *The Journal of Creative Behavior, 2*, 213-214.

Mednick, S. A., & Mednick, M. T. (1967). *Examiner's Manual, Remote Associates Test: College and Adult Forms 1 and 2*. Boston: Houghton Mifflin.

Metcalfe, J. (1986). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(4), 623-634.

Mikulincer, M., & Sheffi, E. (2000). Adult attachment style and cognitive reactions to positive affect: A test of mental categorization and creative problem solving. *Motivation and Emotion, 24*(3), 149-174.

R Development Core Team. (2009). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. Available: http://www.R-project.org.

Schooler, J. W., & Melcher, J. (1995). The ineffability of insight. In S. M. Smith & B. T. Ward & R. A. Finke (Eds.), *The creative cognition approach* (pp. 249-268). Cambridge, MA: The MIT Press.

Shames, V. A. (1994). *Is There Such a Thing as Implicit Problem-solving?* Unpublished doctoral dissertation, The University of Arizona.

Smith, S. M., & Blankenship, S. E. (1989). Incubation effects. *Bulletin of the Psychonomic Society, 27*(4), 311-314.

Sternberg, R. J., & Davidson, J. E. (1982). The mind of the puzzler. *Psychology Today, 16*(6), 37-44.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer

APPENDIX

Instructions and solutions to the insight problems

1. ***Coin problem***: A dealer in antique coins got an offer to buy a beautiful bronze coin. The coin had an emperor's head on one side and the date 544 B.C. stamped on the other side. The dealer examined the coin, but instead of buying it, he called the police to arrest the man. What made him realize that the coin was fake? (Adopted from Metcalfe, 1986).
2. *Solution*: In 544 B.C. there was no knowledge of Jesus Christ as he was as yet unborn. A coin from that time thus could not be marked 'B.C'. Most initial false solutions concern whether the date matched the emperor ruling in 544 B.C., whether bronze was already discovered, etc.
3. ***Egg problem***: Using only one 7-minute hourglass and one 11-minute hourglass, how will you be able to time the boiling of an egg for exactly 15 minutes? (Adopted from Sternberg & Davidson, 1982).
4. *Solution:* Start both hourglasses at the same time. When the 7-minute hourglass runs out (and 4 minutes remain on the 11-minutes hourglass), start boiling the egg. After the 4 minutes have elapsed, turn it over the 11-minute hourglass again to obtain a total time of 15 minutes. An egg is customarily put into a pot of water as soon as it commences to boil. To arrive at the correct solution, the fixedness to approach the problem using this strategy must be overcome.
5. ***Triangle problem*** (spatial problem): The triangle of dots in the picture provided here points to the bottom of the page by moving only three dots? (Adopted from Metcalfe, 1986).
6. ***Solution*****:** Dots to be moved are the dots on the bottom left, bottom right and the top. The correct solution requires a mental rotation.

Problem:

Solution: