Logic and **L**ogical **P**hilosophy
Volume 27 (2018), 517–543
DOI: 10.12775/LLP.2017.015

Natalia Żyluk, Mikołaj Michta,
and Mariusz Urbański

YET ANOTHER SHADE OF DEDUCTION
On measuring deductive flexibility and how it may relate to other cognitive abilities

Abstract. The article describes the construction process of *Deductive Flexibility Test* — considered a difficult deductive reasoning measure — and the research on correlations between fluency in difficult deductive reasoning and other cognitive abilities. The main goal in the research was to examine the relations between *Deductive Flexibility Test* scores and results of *Raven's Advanced Progressive Matrices* — fluid intelligence test. Additionally, the measures of the need for cognitive closure and epistemological understanding were included in the study. The results of the study revealed that *Deductive Flexibility Test* is a reliable instrument and thus can be used for research purposes. We found low or even no statistically significant correlations between the chosen variables. The directions of further research are discussed.

Keywords: deductive reasoning; deductive flexibility; *Deductive Flexibility Test*; intelligence; fluid intelligence; need for cognitive closure; epistemological understanding

1. Introduction

The aim of this article is to describe the construction process of *Deductive Flexibility Test* (DFT) and to present the results of research on correlations between DFT scores and other cognitive abilities, which can be considered as relevant for solving difficult deductive reasoning tasks that DFT consists of. The main goal in this research was to test if there are any significant correlations between the cognitive abilities operationalised by means of results in *Raven's Advanced Progressive Ma-*

Special Issue: Logic, Cognition, Argumentation.
Guest Editors: Mariusz Urbański, Michiel van Lambalgen, and Marcin Koszowy
 Published online May 24, 2017

trices (APM) and *Deductive Flexibility Test* in young adults. Additional goal was to examine correlations between DFT scores and other high-level cognitive abilities such as need for cognitive closure and levels of epistemological understanding.

The article consists of two main parts. In the first one, the process of designing and developing *Deductive Flexibility Test* is presented. We start with introducing the concept of deductive flexibility (Section 2) and provide detailed description of all the stages of test construction, starting with creating its first, pilot version (Section 3), through conducting a pilot study, to developing the final DFT version (Section 4).

The second part is focused on research on correlations between *Deductive Flexibility Test* scores, fluid intelligence measure and measures of some other cognitive abilities (Section 5). The results of the research presented here confirmed that DFT is a reliable tool that can be used along with other instruments. Importantly, the observed correlation patterns can be seen as a starting point for further research focused on deductive reasoning skills and their correlates (Section 6).

2. Deductive flexibility

In many typical verbal deductive reasoning tests subjects are asked to assess, whether certain conclusion can be inferred given presented set of premises (e.g. [8, 9, 18, 26, 28]). Imagine then a task in which conclusion is fixed and you are asked to indicate sets of premises which logically entail that conclusion and — what is important — you can do so only by using predefined information allowed by test designers. Solving this kind of problem requires deductive reasoning skills as well as specific kind of "flexibility", understood here as the ability to switch between different sets of premises that potentially entail a given conclusion and to think about multiple sets of premises simultaneously. In order to grasp the cognitive characteristics highlighted above, Żyluk and Urbański proposed the notion of deductive flexibility [31]. The name of the construct was coined in analogy to cognitive flexibility — an ability to switch between thinking about different concepts and to think about multiple concepts simultaneously [25]. Decision to lean on that concept, while creating a name for a new one, was based on the observation that the ability described by Żyluk and Urbański requires similar, yet more domain — restricted, cognitive skills.

Deductive flexibility can manifest itself in the ability of determining sets of premises that imply a certain conclusion. The phrase "can manifest" is used here, because, although deductive flexibility could easily be characterized in logical terms (referring to relation of logical entailment), its psychological operationalization — in terms of more expanded list of manifestations — requires further analysis.

3. The tool construction

The instrument for measuring deductive flexibility — *Deductive Flexbility Test* — has been developed by Żyluk and Urbański [31]. We describe the general idea underlying the test items' construction, the process of selection and validation of items for a pilot DFT version, the test instructions' development process and the scoring method.

3.1. Test items format

3.1.1. Classical categorical sentences

As a logical framework for DFT Aristotelian logic of classical categorical sentences was chosen (see [2] or [4]). These sentences have administrable and uncomplicated semantics and, as it was shown before, they allow for a straightforward operationalization of easy and difficult deduction tasks [28].

Classical categorical sentences can be divided into four types (S and P are variables denoting classes of objects):

- universal affirmative (All S are P, written as SaP);
- universal negative (No S are P, written as SeP);
- particular affirmative (Some S are P, written as SiP);
- particular negative (Some S are not P, written as SoP).

The intended semantics of these sentences was the one of Aristotle's logic. Thus non-emptiness of terms extensions was assumed, as well as logical interpretation of "some" as opposed to pragmatic one, warranting entailment of particular sentences by universal ones [15].

3.1.2. Single item format

In order to devise correct solutions to the test items operations of conversion, inversion, obversion, contraposition and rules of the square of opposition as well as valid syllogisms were employed [4].

Building blocks of each of the test items are six sentences — five placed above the line, which serve as possible premises, and one under the line, which is a fixed conclusion. Possible premises are selected in such a way that using some combinations of them and the conclusion one can build a valid schema of inference; we shall refer to such combinations as correct ones. However, we allow for a possibility that no combination of premises is correct. A single sentence from the set of possible premises may be used in the different combinations. Importantly, as a correct set of premises is considered only a set without redundant elements, that is, a set that is sufficient to entail the conclusion.

Authors of the tool prepared 93 test items with different types of sentences as conclusions, employing also two types of negation: term (') and propositional ($\neg$) ones.

Structures of exemplary test items (for each of the four types of conclusion) are as following:
Universal affirmative conclusion:

$$\begin{array}{ll} (1) & P'aS' \\ (2) & MaP \\ (3) & SaM \\ (4) & \neg\, SoP \\ (5) & \neg\, SeP \\ \hline & SaP \end{array} \qquad \text{(i)}$$

The solution consists of three combinations of premises: 1 (as the conclusion is equivalent to its contraposition); 2+3 (these two premises with the conclusion form a syllogistic mode *Barbara*) and 4 (given the rules of the square of opposition).
Universal negative conclusion:

$$\begin{array}{ll} (1) & \neg\, PiS \\ (2) & SaM \\ (3) & PeS \\ (4) & MeP \\ (5) & \neg\, SiP \\ \hline & SeP \end{array} \qquad \text{(ii)}$$

The solution consists of four combinations of premises: 1 (given the rules of the square of opposition and operation of conversion), 3 (according to rules of the operation of conversion), 5 (given the rules of the square

of opposition) and $2 + 4$ (these two premises with the conclusion form a syllogistic mode *Celarent*).
Particular affirmative conclusion:

$$\begin{array}{ll} (1) & PaM \\ (2) & MaS \\ (3) & MaP \\ (4) & PiM \\ (5) & \underline{MiS} \\ & SiP \end{array} \qquad \text{(iii)}$$

The solution consists of four combinations of premises: $1 + 2$ (these two premises with the conclusion form a syllogistic mode *Bramantip*); $3 + 2$ (these two premises with the conclusion form a syllogistic mode *Darapti*), $4+2$ (these two premises with the conclusion form a syllogistic mode *Dimaris*) and $3 + 5$ (these two premises with the conclusion form a syllogistic mode *Datisi*).
Particular negative conclusion:

$$\begin{array}{ll} (1) & SaP \\ (2) & P'oS' \\ (3) & PiS \\ (4) & SaM \\ (5) & \underline{PiM} \\ & SoP \end{array} \qquad \text{(iv)}$$

The solution consists of only one premise — 2 (according to rules of the operation of contrapositon).

3.2. Preliminary selection of test items for a pilot version of the tool

Among generated test items, 10 were chosen to be included in a pilot version of the tool. Authors chose these items using the following criteria:

- all four types of conclusions should be used;
- items should vary in terms of number of correct sets of premises and at least one item without correct set of premises should be included;
- the premises should vary with respect to types of negations involved;
- items with propositional negation in conclusion should be used;
- items should vary with respect to types of inferential relations holding between premises and conclusions.

Decision to include 10 items in a pilot version of the tool was based on an assumption, that using more items could cause fatigue and neg-

atively influence participants' motivation to finish the test, as well as their capacity to solve the tasks.

Chosen items were in fact schemas of the final tasks — in the next step variables denoting classes of objects (S, M, P) were replaced with pseudowords (pseudowords were generated as described in [18]) and symbols a, o, e, i with corresponding expressions in Polish. Pseudowords were employed, on the one hand, instead of the real words in order to suppress possible confirmation bias — tendency to favor the information that confirms one's beliefs or views on the world [19]. On the other hand, using pseudowords and not letter variables prevents replacement of the letters with random expressions chosen by a subject (a strategy which may also cause confirmation bias).

To give the Reader an insight into how test items would look like in English[1], we can show a following example. It corresponds to the scheme (iv) above, employing particular negative conclusion:

(1) Every chig is a dazzla
(2) Some non-dazzlas are not non-ozacks
(3) Some dazzlas are chigs
(4) Every chig is an ozack
(5) Some dazzlas are ozacks

Some chigs are not dazzlas

During preparation of final items, decision to use the expressions "some are" and "some are not" in particular affirmative and negative sentences was made. In order to avoid confusion emerging from using pragmatic interpretation (as "Only some are" or "Only some are not"; see e.g. [23]), authors considered using expression "At least one of" as such expression corresponds best with the meaning of "Some are" in the logical interpretation [18]. However, using this expression caused another problem: created sentences turned out to be overcomplicated (consider for example sentence "It is not the case, that at least one of chigs is not an ozack."). Therefore authors decided to use expressions "some are" or "some are not". In order to minimise errors caused by relying on pragmatic interpretation, the authors decided to include a relevant explanation [17]: the information on how to interpret expressions "some are" and "some are not" was added to the instructions.

[1] English version of DFT can also be found on the Reasoning Research Group website: http://reasoning.edu.pl/ (section: Research).

3.3. Competent judges method — final validation of items selected for the pilot version of the tool

Selected 10 test items and preliminary version of instructions for participants were subject to evaluation. As a method of evaluation the authors chose a two-step variant of the so called "competent judges method".

The first group of judges consisted of two experts in logic — a cognitive science PhD student and a logician with a PhD. The judges were asked to solve chosen 10 tasks and to assess, if these items are sufficiently diversified, difficult and adequately built (whether they are presented in the right order in the tool as a whole and whether premises within one test item are ordered in an acceptable way; "rightness" and "acceptability" should be understood in a very subjective manner here, as the judges' verdicts were made on the basis of their own intuitions and not on normative standards of any kind). Among the remarks they conveyed were issues concerning the order of possible premises. Test items have been changed according to judges' suggestions. At this stage, the judges were given only a preliminary version of the test instructions, nevertheless, they submitted comments on its possible content.

In the next step, 10 test items (modified given previous remarks), along with developed test instructions, were presented to a group of third year cognitive science students of Adam Mickiewicz University in Poznań (AMU) ($n = 7$), experienced in logic (cognitive science curriculum at AMU offers several compulsory courses in logic) — in particular well acquainted with the basics of syllogistics. The students' task was to solve the test and to assess, whether the instructions are formulated clear enough and whether test items and the tool as a whole are built properly. Given their remarks, the authors changed the order of the test items, made additional changes concerning the order of premises within selected tasks, added some linguistic corrections and significantly modified the content of the test instruction (short-cutting it, in particular). With a help of the judges, it was possible to estimate the approximate time needed to complete the test. Obtained solutions were instrumental in developing the final scoring method (described in the subsection 3.5).

3.4. Developing final tool instructions

The content of DFT instructions was formulated in such a way as not to overwhelm a participant with the amount of information, while being

sufficiently clearly formulated. Firstly, the authors of the tool had to take care of justifying the use of pseudowords in the test — therefore, while reading the instructions, participants were asked to imagine that they are reading a book whose characters are fantastic creatures living on the planet XYZ, such as chigs or dazzlas. Participants were informed that DFT test items would concern the relations between these creatures.

In the instructions, terms such as "entailment", "deductive", "premises" and "conclusion" were not used; the task was presented more descriptively. The participants were asked to determine, which combinations of sentences from above the line would have to be true in order to accept the sentence under the line. The subjects' task was to indicate as many combinations as possible. Participants were informed, that a single combination may comprise one sentence or more, and that it may be the case that no such combination exists. It was also stressed that the task is to select sentence or sentences that suffice to accept the sentence placed under the line, that is, a combination which does not contain redundant sentences, and that a single sentence can be used in different combinations.

In the instructions an exemplary task with the correct solutions and detailed explanation was presented; it also contained guidelines concerning interpretations of the meaning of "some" (as mentioned previously) and term negations. Additionally, participants were allowed to make notes and to draw pictures on the test materials.

3.5. Scoring method

The solution of a single test item was evaluated taking into account: the total number of responses given by the subject for that item, the number of subject's correct answers (number of indicated correct combinations of premises) for that item and the number of all correct answers for that item. It should be emphasized, that non-indication of any set of premises was also considered as a response (mostly due to the fact that for one of the tasks included in the DFT the correct solution was non-indication of any combination of premises).

Let Z be the set of all possible premises in the single DFT task (there are always five of them). Let us now denote by R the set of all correct solutions (combinations of premises) to a task, and by R_p the set of all the solutions (correct or not) indicated by a subject. Both R and R_p are subsets of the power set of Z ($R \subseteq \mathcal{P}(Z)$; $R_p \subseteq \mathcal{P}(Z)$). When a subject

pointed as a solution the following combinations of premises: 1, $2+3$ and $4+5$, $R_p = \{\{1\},\{2,3\},\{4,5\}\}$; when he or she did not indicate any combination, the set of his or her answers was not empty, but it was a singleton with the empty set as the only element: $R_p = \{\emptyset\}$.

Let us designate the cardinality of a set X (the number of elements in X) by $|X|$. The correctness rate Ev for a single test item for a single subject is calculated according to the following formula:

$$Ev = \frac{2 \cdot |R \cap R_p|}{|R| + |R_p|}$$

Thus, for any single item it is possible to obtain the score ranging from 0 to 1. When the subject's answer contains all and only correct sets of premises ($R = R_p$), it is the case that $Ev = 1$.

As it was stated before, measured ability manifests itself in generating and evaluating non-redundant sets of premises which entail a given conclusion. Because of that, scoring formula promotes giving all correct answers, while lowering score for subjects who give only some correct answers (correct inferences but weaker generation) or give too many answers: both correct and incorrect (vigorous generation but redundant or incorrect inferences). As an example, let us consider possible (not necessarily correct) answers to tasks structured as schemas (i)–(iv), presented in the subsection 3.1:

- in the task structured like (i) (three correct sets of premises; $R = \{\{1\},\{2,3\},\{4\}\}$):
  - the participant A pointed out four sets ($R_p = \{\{1\},\{2,3\},\{4\},\{5\}\}$), of which three were correct, in this case: $Ev = .86$ (approximately);
  - the participant B pointed out two sets ($R_p = \{\{1,2\},\{5\}\}$) — all incorrect — therefore: $Ev = 0$;
- in the task structured like (ii) (four correct sets of premises; $R = \{\{1\},\{3\},\{5\},\{2,4\}\}$):
  - the participant A pointed out one combination of premises: $2+4$ ($R_p = \{\{2,4\}\}$), which was a correct set, in this case: $Ev = .4$;
  - the participant B pointed out three sets of premises: 1, $2+5$, $3+4$ ($R_p = \{\{1\},\{2,5\},\{3,4\}\}$; only one of his answers was correct, in this case: $Ev = .29$ (approximately);
- in the task structured like (iv) (one correct set of premises; $R = \{\{2\}\}$):

- the participant A pointed out one combination of premises ($R_p = \{\{2\}\}$), which was the correct one, thus $Ev = 1$;
- participant B pointed out two combinations of premises ($R_p = \{\{2\}, \{1 + 5\}\}$), of which one was correct, it this case: $Ev = .67$ (approximately).

The total DFT score for a subject was calculated as the average of all obtained item scores. Therefore, the total score also ranged from 0 to 1.

3.5.1. Typology of errors

What deserves a characterization is typology of errors made by the subjects. The possible errors are the following:

- omissions (non-indication of correct sets);
- redundant solutions (correct set, usually having a form of a single sentence, combined with additional sentence or sentences):
  - redundant negative (correct set combined with an incorrect set);
  - redundant positive (correct sets combined);
- completely incorrect solutions.

It should be noted that in the case of redundant solutions (both positive and negative) logical entailment holds between premises and conclusion (as logical entailment in Aristotelian logic is monotonic). In the case of completely incorrect solutions premises do not entail conclusion.

Scoring formula in its present form does not distinguish between redundant solutions and completely incorrect solutions and evaluates them as equally incorrect. As DFT is used to assess subjects' deductive flexibility — understood as an ability manifesting itself in determining non-redundant sets of premises that entail a certain conclusion — such approach is valid. Nevertheless, defining scoring method able to distinguish different types of mistakes, or identifying each type of error as a separate variable category may allow for a deeper analysis of processes involved in solving DFT tasks.

4. The pilot study

The test was administered, in Polish, to 26 participants, students of different curricula with ages ranging from 20 to 25 ($M = 22.65$, $SD = 1.441$). The gender proportion was balanced, with 14 females and 12 males ($\chi^2 = 0.154$; $p > .05$).

The data-set was collected by third-year AMU cognitive science students, as a part of "Empirical Seminar" classes.

item	difficulty index
1.	42.27%
2.	63.52%
3.	36.85%
4.	11.53%
5.	64.83%
6.	36.94%
7.	76.66%
8.	67.17%
9.	53.53%
10.	33.27%
Mean	48.96%

Table 2. Test items difficulty in the pilot version of DFT

4.1. Results

4.1.1. Descriptive statistics

All the statistical analyses were carried out using the statistical software SPSS v. 23.

Table 1 sums up basic descriptive statistics calculated for scores from DFT pilot version. Given the small sample size, to assess the normality of test scores' distribution, the Shapiro-Wilk test was used (last two columns of Table 1). The obtained scores have a normal distribution ($p > .05$).

	M	*SD*	*Me*	*Min*	*Max*	*W*	*p*
DFT	.49	.20	.46	.07	.88	0.977	.794

Table 1. Descriptive statistics for scores obtained in pilot version of DFT and results of S-W test

4.1.2. Difficulty and reliability of the tool

It was assumed that DFT score serves as an indicator of one, homogeneous cognitive ability — deductive flexibility. However, the difficulty of test items varies. Table 2 summarizes difficulty indexes obtained for all 10 DFT test items. The difficulty index for a single item was calculated as an average value of its correct solution rate.

The general recommendation concerning test items' difficulty levels is that they should differentiate between participants on the whole con-

tinuum of the ability these items were designed to measure. Given that, the best solution would be to choose the positions with difficulty index around 50% [11]. DFT items, however, are highly correlated with each other, which excludes application of this principle. As Anastasi and Urbina recommend [1], in such cases the average difficulty index should oscillate around 50%, while single tasks should exhibit varied difficulty — from the hardest to the easiest. In the case of the pilot version of DFT, it was possible to meet both conditions, i.e., the average difficulty oscillates around a desired value and the tasks are diverse in terms of difficulty (see Table 2). What is worth noting, the most difficult test item was item no. 4, in which no combination of premises entails given conclusion.

As DFT consisted of items of different difficulty levels, to assess its reliability, Guttman's λ_2 was used [10]. The obtained reliability level was $\lambda_2 = .893$, which is a sign of very good internal consistency of the tool. Moreover, factor analysis was performed, using Cattel's criterion [5], to confirm that there is only one factor in DFT.

4.2. Developing final version of the tool

As stated previously, authors presumed that DFT should consist of maximum 10 items. During the pilot study it was observed that the 10 item version was still too tiresome (participants were complaining about its length; in some of the cases they decided not to finish the test). In order to provide a tool suitable for use with other questionnaires or tasks in the course of one research, the authors decided to shorten the pilot version of DFT by removing 2 test items.

Decision on which test items were to be removed was not made in consideration of raising reliability level, as that could lead to construction of the tool in which test items would become redundant. Instead of that, the effort was put into maintaining mean level of difficulty around 50% while keeping particular test items difficulties diverse.

Authors removed test item no. 1 (5 correct sets of premises; universal affirmative conclusion) and no. 10. (6 correct sets of premises; universal negative conclusion). In case of test item no. 10, it was fairly easy to find some correct answer but quite difficult to indicate all of them (as almost any combination of two out of five premises was a correct one) and all correct sets of premises would either form a syllogistic mode or be valid ones according to rules of the square of opposition. It is worth reporting, that this item was included in the pilot study version because it has more

correct sets of premises than other test items. What is more, difficulty of item no. 10 in pilot study was similar to difficulty indexes obtained for a few other test items, so its removal did not influence difficulties variety. Test item no. 1 was removed due to the fact, that it correlated weakly with principal component in the factor analysis and it represented low variety of inferences (only one syllogistic mode (*Barbara*) and rules of the square of opposition were employed while creating that item). After the removal, mean (hypothetical) difficulty of 8 items long DFT reached the level of 51.25%.

5. The research: *Deductive Flexibility Test* vs *Raven's Advanced Progressive Matrices* and other cognitive abilities measures

The final version of DFT was administered in the course of research on correlations between fluency in difficult deductive reasoning and fluid intelligence [21] and also some other high-level cognitive abilities: need for cognitive closure [14, 12] and level of epistemological understanding [13, 32].

The majority of data used in the study was collected by the students and the first author of this article.

In the next few subsections the underlying idea and results of this research are briefly described. Given the topic of the paper, the special emphasis is placed on the role of DFT in the course of the whole study. Measures other than DFT scores are discussed only in the context of its relations with DFT results.

5.1. Theoretical motivations

5.1.1. Fluid intelligence

According to Cattell-Horn-Carroll (CHC) model of intelligence, deductive reasoning is identified as one of the reasoning abilities defining fluid factor of general intelligence [7, 24].

Previous research [28] demonstrated, that fluid intelligence and fluency in deductive reasoning are correlated, as predicted by the CHC model of intelligence; however, this holds only for simple, or easy, deductions, as operationalised by syllogistic reasoning. In case of more complex deductions, operationalised, for example, by erotetic inferences (that is, inferences involving questions either as conclusions or as both

premises and conclusions [30]) and polisyllogisms, subjects educational experience needs to be taken into account [28].

Deductive Flexibility Test measures deductive reasoning skills, as solving the test substantially relies on an ability to identify the logical entailment relation. Specifically, given the test complexity, DFT scores can be seen as difficult deduction measures, but of a slightly different type than the one used in the study by Urbański et al. [28].

The main aim of the current research was to test whether there are any significant correlations between cognitive abilities operationalised by means of results in *Raven's Advanced Progressive Matrices* and fluency in difficult deduction tasks measured with *Deductive Flexibility Test*.

Additionally, as correctness of DFT solution requires substantial self-regulative ability, in particular with respect to metalogical norms [22], we were also interested in possible correlations between DFT results and results in tools measuring other high-level cognitive constructs: need for cognitive closure and levels of epistemological understanding.

5.1.2. Need for cognitive closure

As defined by Kruglanski and Webster [14], the need for cognitive closure refers to "individuals' desire for a firm answer to a question and an aversion toward ambiguity" [14, p. 264]. A high level of the need for cognitive closure is associated with preference for order and predictability, ease in decision-making, close–mindedness and discomfort with ambiguity. Individuals characterized by high level of the need for cognitive closure tend to analyse incoming information in a superficial manner and seek for simplified information. On the other hand, a low level of the need for cognitive closure is linked to the higher tolerance for experiencing uncertainly, facilitates the careful and thorough analysis of the situation and fosters the openness to new information [14, 12].

Solving DFT tasks requires an in–depth analysis of the task, switching between different, potentially correct, sets of premises and, in the case of some tasks, in order to obtain the highest score participant has to point more than one correct answer (set of premises). People exhibiting high level of the need for cognitive closure often limit their information processing activities and analyse information in a superficial manner in order to terminate a state in which they feel uncertain as quickly as possible. Given such a characteristics, high level of the need for cognitive closure could lead to different types of errors which may occur during solving DFT test, which, in turn, may be reflected in lower DFT score.

For example, high level of the need for cognitive closure may promote tendency to make omission errors, as such subjects have a tendency for superficial analysis of information, and would be satisfied with some plausible and not necessary the best answer [29] — which would reflect in indicating all the correct sets of premises. Analysing the information in a superficial manner may also facilitate incorrect reactions in the task without any correct set of premises. A tendency for being satisfied with any plausible answer combined with tendencies for pursuit of simplified information and for superficial analysis of information may result in indicating "half–answers" (that is, for example, providing one of a combination of two correct premises for one fixed conclusion), considered in DFT as the completely incorrect solution.

5.1.3. Levels of epistemological understanding

Levels of epistemological understanding can be understood as views on what knowledge is and how it is evaluated and acquired [13, p. 309]. These individual conceptions on knowledge and knowing change during one's lifetime: formation of a mature epistemological understanding is a process that starts with radical objectivism, leading through subjectivism, to an integration of both dimensions. Kuhn et al. [13] distinguished between three levels of epistemological understanding corresponding to these stages: absolutist, multiplist and evaluativist, which are domain-dependent. To this day there is no research on relations between epistemological understanding and reasoning abilities of any kind. As — according to Kuhn and colleagues — the cognitive and intellectual functioning of an individual is significantly determined by his or her views on what knowledge is and how it is evaluated and acquired, the decision of including the variable in the research was made. However, due to a purely exploratory nature of these analyses, no particular hypotheses were proposed.

5.2. Materials and Methods

5.2.1. Participants

The study included 47 participants: students of different curricula with ages ranging from 20 to 25 ($M = 22.72$, $SD = 1.246$). The gender proportion was balanced, with 26 females and 21 males ($\chi^2 = .532$; $p > .05$).

All participants were at least third-year students. We aimed at an age-homogeneous group (as the fluid intelligence and levels of episte-

mological understanding are developmental characteristics [6, 13]), at a similar stage of education. Participants were recruited mainly through social media, some of them volunteered to participate after getting a recommendation from our former participants. Participants did not receive any remuneration for taking part in the study.

5.2.2. Instruments

Aside from *Deductive Flexibility Test*, the data collection tools used in this study included *Raven's Advanced Progressive Matrices* (APM), *Need For Cognitive Closure Scale* (NFCS, shortened version) and *Standardized Epistemological Understanding Assessment* (SEUA). As DFT was described in details previously, its characteristics will be omitted in this part of the paper. All tests were carried out in Polish.

Raven's Advanced Progressive Matrices

Raven's Progressive Matrices results are considered a good measure of reasoning ability component of general intelligence, especially its fluid factor [3]. In the research advanced version of the tool was used. APM contains 48 items, presented as one set of 12 (which is a training set), and another of 36. In each item, the subject is asked to identify the missing element that completes a pattern. Each item is a matrix of figural stimuli that are organized according to latent rules. A subject's aim is to discover proper rules and to choose one correct answer from eight possible ones. An exemplary task (test item in the *Raven's Progressive Matrices* test style) was presented in Figure 1. Items become increasingly difficult as progress is made through each set. Each correct choice is awarded with 1 point (maximum 36 points for the whole APM). If untimed, this test is designed to differentiate between people at the high end of intellectual ability. When administered with time limit, the APM can also be used to assess intellectual efficiency [21]. In this research authors administered APM under time conditions — 40 minutes.

Need For Cognitive Closure Scale

To measure need for cognitive closure authors of the research used shortened, 15 item version of NFCS scale (as there exists a Polish adaptation of this version, described in [12]), with five subscales: desire for predictability, preference for order and structure, discomfort with ambiguity, close-mindedness and decisiveness. NFCS test items have a form of statements expressing possible beliefs, attitudes or experiences. For each

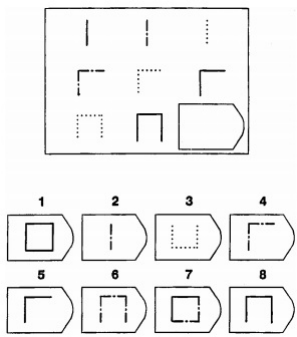

Figure 1. Exemplary matrix in the *Raven's Progressive Matrices* test style [20]

item, a participant is asked to choose one answer that indicates the level to which he or she agrees the statement describes his or hers mindsets, using six-point scale, from 1 = strongly disagree to 6 = strongly agree (the sample item: "I don't like to go into a situation without knowing what I can expect from it."). The higher NFCS score one receives, the higher level of need of cognitive closure he or she exhibits. In the shortened version of the tool it is possible to obtain from 15 to 90 points for the whole instrument (from 3 to 18 within one subscale). In the research both general score and subscales' scores were analysed.

Standardized Epistemological Understanding Assessment

As the fourth instrument *Standardized Epistemological Understanding Assessment* in a paper-and-pencil version was employed [32]. SEUA is a result of a Polish adaptation and modification of an instrument developed by Kuhn, Cheney and Weinstock [13]. The instrument consists of 25 items (5 for each judgement domain) that have a form of pairs of sentences. Each of the items has a form of two incompatible beliefs of two people (for example: "Robin thinks the first painting they look at is better. Chris thinks the second painting they look at is better."). To assess if the transition from the absolutist to the multiplist level has occurred, for each pair of sentences a subject is asked whether only one of presented views could be right, or whether both could have some rightness. The diagnostic answer for the absolutist level was the first one. If the answer is "Both could have some rightness", a subject is asked whether one view could have more merit than another. Negative

answer to that question is considered as a diagnostic for the multiplist level, while affirmative — as a diagnostic for the evaluativist level.

Original version of the tool allows participant to obtain only qualitative scores. Authors of SEUA introduced quantitative scoring method — one participant can obtain 5 scores that reflect the level of his epistemological understanding in each domain (there is no general score).

In our previous research SEUA was administered in an interview setting. Here for the first time SEUA was administered in a paper-and-pencil form. Thus, in addition to conduct correlational analysis using SEUA, the aim of the research was also to check psychometric properties of this tool in a new setting.

5.2.3. Procedure

Tests were administered in two ca. 50 min. sessions, separated by a 5 min. break. During the first session participants were filling APM and NFCS and during the second one, after a break, DFT and SEUA. APM was the only test administered with time limit (40 minutes). Participants were filling the tests individually or in groups up to four people. The study was conducted in the laboratory of the Reasoning Research Group at AMU (http://reasoning.edu.pl/) or, in some cases, outside the laboratory, but under the same standardized conditions.

5.3. Results

5.3.1. Descriptive statistics

All the statistical analysis were carried out using the statistical software SPSS v. 23.

Table 3 presents basic descriptive statistics calculated for all of the variables included in the research and results of the Kolmorogorow-Smirnow test, chosen to assess the normality of scores' distribution.

In the case of SEUA 46 answers were received instead of 47, since SEUA score of a participant who did not understand the SEUA instructions properly was removed from analysis. APM, DFT, general NFCS score and two of its subscales: preference for order and structure and decisiveness, had a normal distribution; all of the SEUA subscales scores and three of NFCS subscales scores did not have a normal distribution (see Table 3 for Z and p-value obtained for K-S test).

instrument	n	M	SD	Me	Min	Max	K-S test	
							Z	p
APM	47	26.34	4.425	25	14	34	0.114	.157
DFT	47	.55	.189	.55	.177	.975	0.063	.200
NFCS: general	47	52.32	6.985	52	29	63	0.115	.154
NFCS: desire for predictability	47	9.87	2.708	10	4	16	0.136	.030
NFCS: preference for order and structure	47	11.47	3.161	11	3	18	0.114	.160
NFCS: discomfort with ambiguity	47	13.53	2.977	14	4	18	0.222	<.001
NFCS: close-mindedness	47	7.87	1.907	8	3	12	0.191	<.001
NFCS: decisiveness	47	9.53	3.028	9	3	15	0.112	.185
SEUA: personal taste judgements	46	10.54	1.394	10	8	14	0.196	<.001
SEUA: aesthetics judgements	46	10.59	1.147	10	10	14	0.413	<.001
SEUA: value judgements	46	11.15	2.582	11	5	15	0.154	.008
SEUA: truth judgements: social world	46	13.09	2.009	13	5	15	0.200	<.001
SEUA: truth judgements: physical world	46	11.96	2.590	12.5	5	15	0.156	.006

Table 3. Descriptive statistics for all variables included in the study and results of K-S test

Item	Difficulty index
1.	74.46%
2.	38.01%
3.	8.51%
4.	72.26%
5.	47.71%
6.	74.26%
7.	75.07%
8.	53.38%
mean:	55.46%

Table 4. Test items difficulty in final version of DFT

5.3.2. Difficulty analysis

For the DFT items difficulty level was calculated in the same way as it was done for its pilot version. In the case of the final DFT version, the pattern of difficulty analysis results were similar to the pilot DFT version (see Table 4.). As final DFT version was a result of deleting items no. 1 and 10, the previous item number (from pilot version) corresponding to certain item number n (from final version) will be $n-2$. It was observed that, once again, test item for which the only correct solution was the empty set (item no. 3 in final version and item no. 4 in pilot version) was the most difficult for participants. The average difficulty level of the tool was satisfactory.

5.3.3. Reliability analysis

To asses the reliability of the tests' scores Cronbach's α coefficient and Guttman's λ_2 were used. As APM and DFT consist of items of varying difficulty levels, to assess their reliability Guttman's λ_2 was employed. In the case of SEUA, reliability was assessed using Cronbach's α coefficient for each judgment domain. Cronbach's α was also used for assessing the reliability of NFCS general score and NFCS subscales. The obtained reliability levels are summarized in Table 5.

These results are consistent with those reported in previous Polish research (e.g. [12, 28, 31]). One significant exception is SEUA administered for the first time in a paper-and-pencil setting. Analysis showed lower levels of reliability than the ones obtained in an interview setting [32].

instrument	reliability level	
	Cronbach's α	Guttman's λ_2
APM	-	.789
DFT	-	.820
NFCS: general	.626	-
NFCS: desire for predictability	.659	-
NFCS: preference for order and structure	.847	-
NFCS: discomfort with ambiguity	.711	-
NFCS: decisiveness	.763	-
NFCS: close-mindedness	.552	-
SEUA: personal taste judgements	.436	-
SEUA: aesthetics judgements	.764	-
SEUA: value judgements	.647	-
SEUA: truth judgements: social world	.721	-
SEUA: truth judgements: physical world	.670	-

Table 5. Results of reliability analysis for all variables included in the study

5.3.4. Correlational analysis

As it was mentioned before, in this article we focus only on correlations between DFT scores and other variables included in the research, with the particular emphasis on DFT–APM relations, as it was the main point of interest in the research.

To assess the relationships between abilities mentioned above, Pearson product–moment correlation coefficients and Spearman's rank correlation coefficients were computed (the choice of the indicator depended on the scores' distribution normality).

The Pearson product–moment correlation coefficient was computed to assess the relationship between DFT and APM scores and also DFT and NFCS general scores. To determine if DFT scores are correlated with NFCS subscales, both the Pearson product–moment correlation coefficient and Spearman's rank correlation coefficient were used. Spearman's rank correlation coefficient was computed also in order to assess, if DFT are related to SEUA subscales. The results of correlational analysis are summarized in Table 6.

instrument	DFT
APM	$r = .295$*
NFCS: general	$r = -.179$
NFCS: desire for predictability	$r_s = -.208$
NFCS: preference for order and structure	$r = -.107$
NFCS: discomfort with ambiguity	$r_s = -.072$
NFCS: decisiveness	$r = -.040$
NFCS: close-mindedness	$r_s = .073$
SEUA: personal taste judgements	$r_s = .007$
SEUA: aesthetics judgements	$r_s = .096$
SEUA: value judgements	$r_s = .001$
SEUA: truth judgements: social world	$r_s = .322$*
SEUA: truth judgements: physical world	$r_s = .115$

$*p < .05$

Table 6. Correlational analysis results

5.4. Discussion

The purpose of this study was to determine whether there are significant correlations between cognitive abilities operationalised by means of results in *Raven's Advanced Progressive Matrices* and fluency in difficult deduction tasks measured with *Deductive Flexibility Test*. The secondary goal was to test possible correlations between DFT test results and other cognitive abilities: the need for cognitive closure (measured with *Need for Cognitive Closure Scale*) and levels of epistemological understanding (measured with *Standardized Epistemological Understanding Assessment*).

In most of the cases, satisfactory reliability levels were obtained. What is worth noting, *Deductive Flexibility Test* reliability level exceeded .8 threshold (Guttman's $\lambda_2 = .82$). It was observed that APM and DFT scores were significantly correlated. The correlation was low, though. In line with previous studies on correlations between levels of fluid intelligence and two kinds of deduction [28], the obtained results confirm that the concept of deduction proposed by CHC model might not be finegrained enough as to account for many manifestations of deductive reasoning. However, further research are needed in order to provide more comprehensive explanation of such results. In our future research we would like to take into account the impact of logical experience —

as it was done in the study mentioned previously [28] — by inviting to our research the participants who differ in level of experience in learning logic. The main conclusion made by Urbański et al. [28] was that fluency in difficult deductions, while related to fluid intelligence, depends also on subjects' experience and that this does not hold in case of simple deductions. They observed that in the group trained in logic there was no correlation between fluency in difficult deduction and level of fluid intelligence, while in groups that did not have intensive logical training there were moderate to high correlations between these variables. At the same time, results obtained in difficult deduction tasks by the group trained in logic were significantly higher than results of non-trained groups (one with the same fluid intelligence level as trained group and other with significantly lower).

It may also be the case that difficult deductive reasoning is only one of the abilities involved in solving DFT and that can also explain the low correlation. The future research should be aimed at examining the relationship between the DFT scores and other cognitive measures, such as working memory, in order to provide more accurate specification of the skill-set involved in solving DFT. It would also be interesting to examine the relations between the DFT scores and results of other deductive abilities test (e.g. simple/easy deduction test — *Deductive Reasoning Test* or difficult deduction tests — *Deductive Reasoning Styles* and *Erotetic Reasoning Test* [28]). Worth consideration is error analysis or developing scoring method which takes into account different types of errors committed by subjects. Interviews with participants while filling DFT test would also provide valuable information about DFT measures.

In the case of DFT and NFC scores no significant correlations were found. One explanation of such result can be the fact that although people with the high need for cognitive closure often limit their information-processing activities, in some of particular instances in which closure is still lacking after their initial answer-seeking process, their need for closure may promote extensive information processing [14]. This issue may be addressed in further research in the interview setting.

Additionally, it turned out that SEUA test is more suitable for using in the interview setting than in a paper-and-pencil form. During interviews with original version of the tool most of the participants needed support of a researcher in order to understand the content of the instructions and of the test items correctly. In a paper-and-pencil setting it was not possible to react immediately in cases of any misunderstand-

ing, simply because participants rarely realized they had problems with understanding their task. On the basis of comments they provided after research, it is possible to conclude that very often the task was misunderstood. These comments, the fact that one of the participants did not fill the questionnaire in an appropriate way and not very good reliability of the tool were reasons that decision was made not to interpret the results of correlational analyses involving SEUA scores (in current paper-and-pencil setting) and other tools. Such an analysis could be repeated after introducing reliability-increasing corrections to paper-and-pencil version of SEUA.

6. General conclusions and further research

The article presented the process of creating final version of *Deductive Flexibility Test* and described the research on relations between deductive flexibility and other cognitive abilities, fluid intelligence in particular.

Deductive Flexibility Test is reliable and thoroughly designed tool to measure what its authors defined as deductive flexibility. However, it should not be restricted to quantitative assessment of ability it was designed to measure. It appears that the novel structure of test items, where subject's task is to find smallest sets of premises that lead to a given conclusion, makes it possible to use DFT as a tool to assess different strategies used by participants in the course of solving difficult deductive tasks. Analysis of errors and patterns in which subjects choose and test sets of premises could also offer an insight into abductive reasoning processes [16, 27], as solving DFT can reasonably be interpreted as a specific case of abduction. As mentioned before, such research should have a form of a recorded interview. Apart from verbal report, also drawings and notes made by participants can serve as a valuable source of information concerning problem-solving processes engaged in filling DFT. However, further research is needed in order to provide an adequate psychological operationalization of the "deductive flexibility" notion; future research on relations between DFT scores and different aspects of cognitive functioning of an individual may help in addressing this issue.

More detailed analysis concerning abilities involved in solving DFT tasks would also help to interpret correlational patterns between DFT scores and other cognitive processes measures. In particular, such an analysis, followed by appropriately conduced research, would make possi-

ble to differentiate — in terms of relations with fluid intelligence measures and an individual experience — between different types of deductions.

These two lines of research sketched above, can be, in fact, conducted simultaneously, as they both are aimed at providing more "backstage" information concerning our cognitive skills, deductive reasoning in particular.

Acknowledgments. Research reported in this paper were supported by the National Science Centre, Poland (DEC-2013/10/E/HS1/00172).

The authors would like to thank Kinga Antonik-Jonczyk, Dajana Bieganowska, Agnieszka Dubowska, Dominika Koch, Iga Kropa, Maciej Małkowski, Dawid Ratajczyk, Anita Steć and Karolina Karpe for their participation in data collection process.

References

[1] Anastasi, A., and S. Urbina, *Psychological Pesting*, Prentice Hall, Upper Saddle River, NJ, 1997.

[2] Aldrich, H., and J. Hill, *Artis Logicae Rudimenta: With Illustrative Observations on Each Section*, J. Parker, Oxford, 6th ed., Kessinger Publishing reprint from 2007, Whitefish, MT, 1850.

[3] Blair, C., "Fluid cognitive abilities and general intelligence: A lifespan neuroscience perspective", in R. M. Lerner and W. F. Overton (eds.), *The Handbook of Life-Span Development. Cognition, Biology, and Methods*, John Wiley & Sons, Hoboken, NJ, 2010. DOI: 10.1002/9780470880166.hlsd001008

[4] Bocheński, J. M., *A History of Formal Logic*, University of Notre Dame Press, Notre Dame, IN, 1961.

[5] Cattell, R. B., "The scree test for the number of factors", *Multivariate Behavioral Research* 1, 2 (1966): 245–276. DOI: 10.1207/s15327906mbr0102_10

[6] Horn, J. L., and R. B. Cattell, "Age differences in fluid and crystallized intelligence", *Acta Psychologica* 26 (1967): 107–129. DOI: 10.1016/0001-6918(67)90011-X

[7] Flanagan, D., "The Cattell-Horn-Carroll (CHC) theory of cognitive abilities", pages 368–386 in C. Reynolds, K. Vannest, and E. Fletcher-Janzen (eds.), *Encyclopedia of Special Education*, John Wiley and Sons, Hoboken, NJ, 2008. DOI: 10.1002/9780470373699.speced0381

[8] Gilinsky, A. S., and B. B. Judd, "Working memory and bias in reasoning across the life span", *Psychology and Aging* 9, 3 (1994): 356–371. DOI: 10.1037/0882-7974.9.3.356

[9] Goel, V., B. Gold, S. Kapur, and S. Houle, "The seats of reason? An imaging study of deductive and inductive reasoning". *NeuroReport* 8, 5 (1997): 305–1310.

[10] Guttman, L., "A basis for analysing test-retest reliability", *Psychometrika*, 10 (1945): 255–282. DOI: 10.1007/BF02288892

[11] Hornowska, E.. *Testy psychologiczne: Teoria i praktyka* (Psychological tests: Theory and practice), Wydawnictwo Naukowe "Scholar", Warszawa, 2014.

[12] Kossowska, M., K. Hanusz, and M. Trejtowicz, "Skrócona wersja Skali Potrzeby Poznawczego Domknięcia. Dobór pozycji i walidacja skali" (Short version of the Need for Cognitive Closure Scale: Items selection and scale validation), *Psychologia Społeczna* 7, 1 (2012): 89–90.

[13] Kuhn, D., R. Cheney, and M. Weinstock, "The development of epistemological understanding", *Cognitive Development* 15, 3 (2000): 309–328. DOI: 10.1016/S0885-2014(00)00030-7

[14] Kruglanski, A., and D. M. Webster, "Motivated closing of the mind: "Seizing" and "freezing"", *Psychological Review* 103 (1996): 263–283. DOI: 10.1037/0033-295X.103.2.263

[15] Łukasiewicz, J., *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic*, Clarendon Press, Oxford, 1951.

[16] Magnani, L., *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*, Springer, Berlin, 2009.

[17] Newstead, S. E., "Gricean implicatures and syllogistic reasoning", *Journal of Memory and Language* 34, 5 (1995): 644–664. DOI: 10.1006/jmla.1995.1029

[18] Paluszkiewicz, K., "Polisyllogisms — report", Research report, AMU Institute of Psychology, Poznań, 2014.

[19] Plous, S., *The Psychology of Judgment and Decision Making*. Mcgraw-Hill Book Company, New York, 1993.

[20] Raven, J., "The Raven's progressive matrices: Change and stability over culture and time", *Cognitive Psychology*, 41, 1 (2000): 1–48. DOI: 10.1006/cogp.1999.0735

[21] Raven, J., J. C. Raven, and J. H. Court, *Manual for Raven's Progressive Matrices and Vocabulary Scales* (Section 4: "Advanced progressive matrices"), Harcourt Assessment, San Antonio, TX, 2003.

[22] Ricco, R. B., and W. F. Overton, "Dual systems Competence ↔ Procedural processing: A relational developmental systems approach to reasoning", *Developmental Review*, 31, 2 (2011), pp 119–150. DOI: 10.1016/j.dr.2011.07.005

[23] Schmidt, J. R., and V. A. Thompson, " 'At least one' problem with 'some' formal reasoning paradigms", *Memory & Cognition* 36, 1 (2008): 217–229. DOI: 10.3758/MC.36.1.217

[24] Schneider, W., and K. McGrew, "The Cattell-Horn-Carroll model of intelligence", pages 99–144 in D. Flanagan and P. Harrison (eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, Guilford Press, New York, NY 2012.

[25] Scott, W. A., "Cognitive complexity and cognitive flexibility", *Sociometry* 4 (1962): 405–414. DOI: 10.2307/2785779

[26] Shynkaruk, J. M., and V. A. Thompson, "Confidence and accuracy in deductive reasoning", *Memory & Cognition* 34, 3 (2006): 619–632. DOI: 10.3758/BF03193584

[27] Urbański, M., *Rozumowania abdukcyjne* (Abductive Reasoning), Adam Mickiewicz University Press, Poznań, 2009.

[28] Urbański, M., K. Paluszkiewicz, and L. Urbańska, "Deductive reasoning and learning: A cross-curricular study", Research report, AMU Institute of Psychology, Poznań, 2014.

[29] Webster, D. M., and A. W. Kruglanski, "Individual differences in need for cognitive closure", *Journal of Personality and Social Psychology* 67, 6 (1994): 1049–1062. DOI: 10.1037/0022-3514.67.6.1049

[30] Wiśniewski, A., *Questions, Inferences, and Scenarios*, College Publications, London, 2013.

[31] Żyluk, N., "Test giętkości dedukcyjnej — raport z konstrukcji narzędzia" (Deductive Flexibility Test — development report), Research report, AMU Institute of Psychology, Poznań, 2016.

[32] Żyluk, N., K. Karpe, M. Michta, W. Potok, K. Paluszkiewicz, and M. Urbański, "Assessing levels of epistemological understanding: The Standardized epistemological understanding assessment (SEUA)", *Topoi* (2016). DOI: 10.1007/s11245-016-9381-4

NATALIA ŻYLUK, MIKOŁAJ MICHTA, AND MARIUSZ URBAŃSKI
Department of Logic and Cognitive Science
Institute of Psychology
Adam Mickiewicz University in Poznań, Poland
murbansk@amu.edu.pl