British Journal of Educational Psychology (2019), 89, 501–523

The British Psychological Society

www.wileyonlinelibrary.com

Special Issue Article

The ability to evaluate arguments in scientific texts: Measurement, cognitive processes, nomological network, and relevance for academic success at the university

Hannes Münchow[1]*  , Tobias Richter[1], Sarah von der Mühlen[1] and Sebastian Schmid[2]

[1]Department of Psychology IV, University of Würzburg, Germany
[2]Department of Pedagogy, University of Regensburg, Germany

Background. The evaluation of informal arguments is a key component of comprehending scientific texts and scientific literacy.

Aim. The present study examined the nomological network of university students' ability to evaluate informal arguments in scientific texts and the relevance of this ability for academic success.

Sample. A sample of 225 university students from the social and educational sciences participated in the study.

Methods. Judgements of plausibility and the ability to recognize argumentation fallacies were assessed with a novel computer-based diagnostic instrument (Argument Judgement Test; AJT).

Results. The items of the AJT partly conform to a 1-PL model and test scores were systematically related to epistemological beliefs and verbal intelligence. Item-by-item analyses of responses and response times showed that implausible arguments were more difficult to process and correct responses to these items required increased cognitive effort. Finally, the AJT scores predicted academic success at university even if verbal intelligence and grade point average were controlled for.

Conclusion. These findings suggest that the ability to evaluate arguments in scientific texts is an aspect of rationality, relies on reflective processes, and is relevant for academic success.

Scientific discourse is characterized by rational dispute and the exchange of arguments. Therefore, the ability to comprehend and assess the validity of scientific arguments is an important facet of scientific literacy, which may be functionally defined as 'the ability of people to understand and critically evaluate scientific content in order to achieve their goals' (Britt, Richter, & Rouet, 2014, p. 105). The ability to comprehend and evaluate

**Correspondence should be addressed to Hannes Münchow, Department of Psychology IV, University of Würzburg, Röntgenring 10, 97070 Würzburg, Germany (email: hannes.muenchow@uni-wuerzburg.de).*

DOI:10.1111/bjep.12298

arguments may also be regarded as the receptive component of the broader ability to argue scientifically (Osborne, 2010). Research in primary and secondary school classrooms suggests that fostering the ability to argue scientifically might benefit learning scientific concepts (e.g., Adey & Shayer, 1993; Mercer, Dawes, Wegerif, & Sams, 2004). Nevertheless, the ability to argue scientifically and, in particular, knowledge and skills to evaluate arguments in scientific texts are seldom taught explicitly in school, where science often appears to students as 'a monolith of facts' (Osborne, 2010, p. 464) instead of a discursive endeavour based on the exchange of arguments.

Consequentially, it cannot be taken for granted that students entering university master this ability. For example, a study by von der Mühlen, Richter, Schmid, Schmidt, and Berthold (2016) found that first-year psychology students performed worse than psychological scientists in tasks that involve judging the plausibility of informal arguments and detecting argumentation fallacies. This finding suggests that the ability to evaluate arguments in scientific texts might be relevant for academic success at the university. However, overall, research on this issue is surprisingly scarce, partly because it is unclear how the focal ability can be measured. There is the well-established Argument Evaluation Test to assess individual tendencies to fall prey to a belief bias in evaluating informal arguments (Stanovich & West, 1997) but, to our knowledge, no established method exists to measure the ability to evaluate informal arguments embedded in coherent scientific texts.

Arguments and argumentation are characteristic for all scientific disciplines but the contents and the forms of the arguments presented may vary between disciplines. Moreover, the ability to evaluate arguments is likely to depend in part on the familiarity with discipline-specific texts and conceptual knowledge (as a kind of discipline expertise, Rouet, Favart, Britt, & Perfetti, 1997). Therefore, we conceive of this ability as a general (discipline-independent) ability that is nevertheless acquired in the context of specific scientific discipline and requires a discipline-specific approach for assessment.

In this study, we used a novel computer-based method to assess the ability to evaluate arguments in scientific texts in the area of psychology and related behavioural, educational, and social sciences. In particular, we investigated the relationships of this ability to other constructs, such as general cognitive capability, argument comprehension, and epistemological beliefs. The goal was to derive a nomological network and examine whether this ability makes a unique contribution to the prediction of academic success at the university. The guiding theoretical idea of the study was that argument evaluation is a rational activity that only partially depends on intelligence. This general assumption was explored further in detailed analyses of drivers of item difficulty and item-by-item response times.

Evaluating informal arguments

Research on reasoning often focuses on the evaluation of formal, deductive arguments, in which a conclusion follows with logical necessity from a set of premises. By contrast, scientific texts usually contain informal arguments. In such arguments, the claim is not necessarily true but more or less probable given the supportive reasons presented in the argument (Green, 1994; Voss & Means, 1991). Moreover, informal arguments are expressed in natural language and can be couched in many different linguistic structures, which makes it more difficult to comprehend such arguments and, presumably, more difficult to evaluate their plausibility. For distinguishing strong from weak informal arguments, not only the accuracy of the presented information, but also the completeness and the internal consistency of the arguments need to be considered (Shaw, 1996). Weak informal arguments provide implausible information because relevant linkages between

reasons and conclusion are either missing or very unlikely. Argumentation fallacies, such overgeneralization or circularity, clearly lead to weak arguments that violate the criteria of completeness and internal consistency, for example, by providing irrelevant reasons or neglecting information that affects the probability of the conclusion.

Argument evaluation and academic achievement

Informal arguments are the building blocks of scientific texts and can appear in many forms, such as the presentation of evidence for theoretical claims, providing theoretical reasons for specific hypotheses, or explanations for observations. It is a core characteristic of science that scientists are in disagreement about theoretical explanations, and the exchange of arguments in a rational dispute may be regarded as the driver of scientific progress. Consequentially, for making sense of scientific texts, readers constantly need to compare and evaluate arguments within and across texts. The abilities to evaluate informal scientific arguments accurately are key components of scientific literacy (Britt *et al.*, 2014) and highly relevant for academic success. At the same time, studies have shown that university students often have difficulties with normatively adequate evaluation of arguments. In many cases, students base their evaluation solely on the believability of information or intuitive judgements but neglect the criteria of completeness and the internal consistency (e.g., Shaw, 1996; von der Mühlen *et al.*, 2016).

Argument evaluation and cognitive effort

Why is the evaluation of informal arguments so difficult? One answer is that it is cognitively demanding. From the perspective of mental model theory (Johnson-Laird, 1983), judging the internal consistency of informal arguments requires readers' to construct and maintain more than one mental model in working memory. Moreover, readers need to actively search their long-term memory to judge the completeness of the reasons. In particular, judging the plausibility of weak informal arguments requires more cognitive effort compared to making judgements about strong informal arguments. To make a reasoned judgement as to whether an argument is implausible, readers need to double-check whether they have understood the argument correctly and whether they might have overlooked unstated premises that might render the argument plausible.

Argument evaluation, cognitive ability, and rationality

Argument evaluation is an instance of rational thinking and the ability to evaluate arguments is an aspect of epistemic (evidential) rationality, which 'concerns how well beliefs map onto the actual structure of the world' (Toplak, West, & Stanovich, 2013, p. 8). Rationality as an individual disposition may be distinguished conceptually and psychometrically from general cognitive ability as measured by typical intelligence tests (Stanovich, 2012; Stanovich & West, 1998). While verbal intelligence, for example, is certainly helpful for processing arguments, the evaluation of arguments requires skills that go beyond those required for verbal intelligence tasks. Moreover, being an aspect of epistemic rationality, the ability to evaluate arguments should be related to epistemological beliefs, that is, a person's beliefs about the nature of knowledge and knowing (Hofer & Pintrich, 2002). Epistemological beliefs are relevant for learning with texts in higher education (e.g., Richter & Schmid, 2010; Rosman, Peter, Mayer, & Krampen, 2018). Specifically, the ability to evaluate arguments in scientific texts might be related to beliefs

about the structure and changeability of scientific knowledge. Only students who believe that scientific knowledge is structured but also changeable (as opposed to representing eternal truths) may engage in the cognitively effortful activity of argument evaluation.

Research questions and hypotheses

We used a novel computer-based diagnostic instrument to assess university students' ability to evaluate arguments in scientific texts in order to explore the nomological network of the construct, its relevance for academic success, and the underlying cognitive processes. We examined the psychometric properties of this instrument, especially whether it conforms to the Rasch model, and used the test scores for addressing the following research questions and hypotheses.

We assumed the evaluation of implausible arguments to be cognitively more demanding than that of plausible arguments, with the consequence that these arguments should be more difficult than plausible items (Hypothesis 1). Moreover, response times should be longer in arguments correctly identified as implausible, indicating that more cognitive effort is needed to evaluate weak arguments correctly (Hypothesis 2). As an exploratory research question, we also examined the role of information density: arguments containing more information, that is, making more information explicit, might be easier to evaluate than arguments that are more elliptic.

With regard to the nomological network of the ability to evaluate arguments in scientific texts, we expected positive relationships of the test scores with response times (Hypothesis 3), argument comprehension (Hypothesis 4), the epistemological beliefs that scientific knowledge is structured (Hypothesis 6) and changing (Hypothesis 6), and verbal intelligence (Hypothesis 7).

Finally, in analogy with studies showing that the ability to argue scientifically is related to science learning at school (Osborne, 2010), we assumed that the ability to evaluate arguments in scientific texts would be relevant for academic success at the university, which implies that this ability should correlate with students' current average grade (Hypothesis 8). Importantly, this relationship should hold even if verbal intelligence and the grade point average (GPA) obtained in school, two powerful predictors of academic success, are controlled for (Hypothesis 9).

Method

Sample

Two hundred and twenty-five students from two German universities took part in the validation study for the Argument Judgement Test (AJT; 77.3% women, 22.7% men). Participants were recruited from the social and educational sciences and participated voluntarily. However, it was ensured that they had not yet obtained their Bachelor's degree. The sample was a convenience sample. The average age of the students was 23.6 years (SD = 5.4), and the average number of semesters was 3.3 (SD = 2.9). One hundred and forty-two participants (63.1%) studied psychology, 73 (32.4%) teaching professions and 10 (4.4%) other subjects (e.g., social work).

Furthermore, in an independent sample of psychology students (N = 22), the AJT was administered twice with a 13-month interval to determine the instruments' predictive validity. This sample was mostly female (17, 77.3% women; 5, 22.7% man) with an average age of 22.50 years (SD = 4.00), ranging from 19 to 35.

Procedure

Upon arrival in the laboratory, participants were welcomed and informed about the purpose, duration, and procedure of the study and provided informed consent. The study lasted approximately 90 min and was completely computer-based. Participants received 12 Euros or four Euros plus partial course credits.

Participants were tested in groups of up to eight persons. First, they provided demographic information such as age, gender, the grade of their university entrance qualification, and their current grade average. Afterwards, participants worked on a test battery that included the AJT and a test assessing argument comprehension (Argument Structure Test, Münchow, Richter, von der Mühlen, Schmid, Bruns, & Berthold, in press) as well as the participants' epistemological beliefs about psychology as a science. In addition, their verbal intelligence was assessed.

Instruments

All instruments were presented in German. The German items and English translations for the AJT can be found in the Appendix.

Argument judgement test

The AJT consists of two expository texts from psychology, with one text addressing smoking behaviour (550 words; adapted from Fuchs & Schwarzer, 1997; see also Schroeder, Richter, & Hoever, 2008) and the other text addressing objective self-awareness (404 words; adapted from Bierhoff, 1993; Brehm & Kassin; Herkner, 1991). Each text contains 15 short arguments of varying plausibility that contain a claim and at least one reason justifying the claim. Ten arguments in each text are plausible in the sense that the supporting of the claim is strong and reasonable. The remaining five arguments per text are implausible arguments that show poor and defective justification for the claim. Implausibility was created by implementing common argumentation fallacies into the arguments (i.e., contradiction, false dichotomy, wrong example, circular reasoning, overgeneralization; cf. Dauer, 1989).

The AJT is divided into two parts. In Part 1, participants evaluate each argument while reading the text by pressing a key labelled 'plausible' or 'implausible' on a keyboard in front of them. Subsequently, the next part of the text with the next argument appears. For Part 1, two different test scores are computed: (1) the proportions of the arguments correctly evaluated as plausible or implausible, and (2) a d'-test score based on signal detection theory. The d'-score takes individual response tendencies into account by correcting true-positive judgements of implausible arguments as being implausible (hits) by false-positive judgements of plausible judgements of implausible (false alarms; Macmillan & Creelman, 2004). In the AJT, plausible and implausible arguments are not exactly balanced, which is why d' was used as an alternative test score.

In Part 2, participants assign the arguments identified as implausible in Part 1 to a list of five common argumentation fallacies (short explanation sentences were given for each fallacy). Participants can also choose one of the following response options: 'I don't know', 'I was wrong, there is no error', or 'None of the above mentioned errors, but . . .'. For the last option, participants can enter a response into a text box. In Part 2, the proportions of the correctly assigned arguments serve as test score. Note that implausible arguments mistakenly evaluated as plausible in Part 1 are also scored as false responses in Part 2. Consequentially, the test scores derived from Part 1 and Part 2 of the test are not completely independent of each other.

When creating the AJT, care was taken that argumentation fallacies only occurred within one argument in order to avoid global inconsistencies in the text. The argumentation fallacies that were chosen for the implausible arguments in the AJT are widely known and their detection does not demand any formal training in argumentation theory or logic. However, one of the items intended to reflect a plausible argument had to be removed from data analyses because its formulation could have been wrongly interpreted as an argumentation fallacy. Figure 1 shows an example item of the AJT.

Argument structure test

A test of argument comprehension, the Argument Structure Test (AST; Münchow *et al.*, in press), was included as a measure of criterion validity. The AST is a computer-assisted diagnostic tool for measuring the competence to recognize and correctly assign functional argument components. Short informal arguments (M = 104 words, SD = 24 words) are

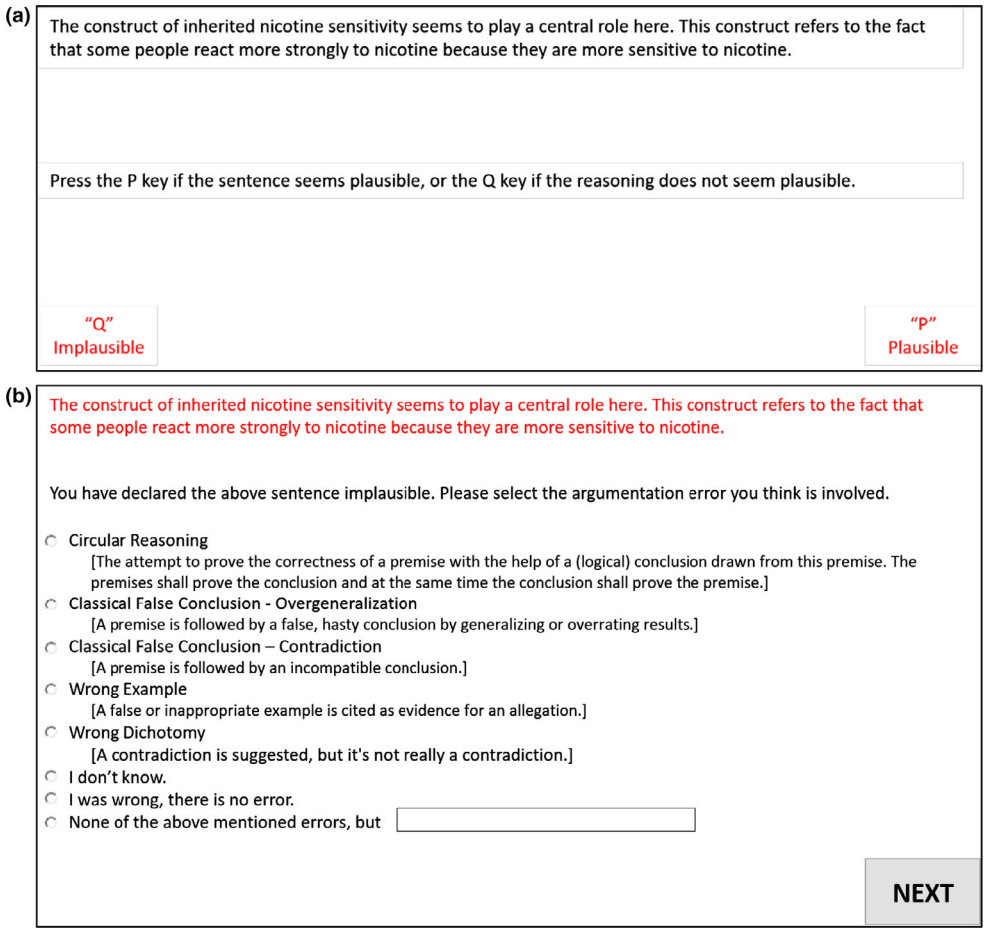

Figure 1. Example item for (a) Part 1 and (b) Part 2 of the AJT. Please note that the items were presented in German in this study. In order to present the general idea of the AJT's procedure more clearly, this example item has been translated into English. [Colour figure can be viewed at wileyonlinelibrary.com]

presented that contain several functional argument components according to Toulmin (1958). Each argument component corresponds to one sentence in an argument. The task is to assign the five argument components for each of the eight arguments (40 items in total). The number of correctly assigned argument components reflects the competence to understand the structure of informal arguments (Cronbach's $\alpha = .76$).

Epistemological beliefs

Epistemological beliefs about psychology as a science were assessed with the CAEB questionnaire (Connotative Aspects of Epistemological Beliefs; Stahl & Bromme, 2007). The CAEB consists of 24 pairs of opposing adjectives (e.g., 'simple' vs. 'complex') that form a semantic differential using a seven-point scale. Two subscores can be formed to assess two dimensions of epistemological beliefs, texture (Cronbach's $\alpha = .72$) and variability of knowledge (Cronbach's $\alpha = .71$).

Verbal intelligence

Verbal intelligence as a measure of cognitive capability was assessed with the subtest sentence completion, verbal analogies and similarities from the basic module of the I-S-T 2000R (Intelligenz-Struktur-Test 2000 R [Intelligence Structure Test 2000R]; Amthauer, Brocke, Liepmann, & Beauducel, 2001), which were aggregated to an index of verbal intelligence (Cronbach's $\alpha = .82$).

Grade point average

We also asked participants for their average grade in their school leaving certificate of the German academic track high school ('Abiturnote'). The grades range from 1 = 'very good' to 5 = 'unsatisfactory'.

Current average grade

As a measure of academic success, participants were asked for their current grade average in their study programme (ranging from 1 = 'very good' to 5 = 'unsatisfactory'). Only 77 participants (34%) provided their current grade average. All of them were students of psychology. Because participants were free to provide this information or not, selection effects regarding the students' reports of their current grades might have occurred. However, AJT scores did not differ between students who reported their current average grades and those who did not report their grades at certain levels of performance in the AJT. According to Enders (2010), this indicates that the type of missingness in this variable was missing at random. Moreover, the majority of the participants were still at the beginning of their studies and probably had not taken any examinations yet. In fact, 96 (42.7%) of the students were in their first semester and 125 (55.6%) students were still within their first two semesters of studies.

Soft- and hardware

All tests were administered via desktop computers with 24″-computer screens. Responses and response times for each input (computer mouse, keyboard) were recorded.

Results

Item and scale characteristics of the AJT

The internal consistencies (Cronbach's α) for Part 1 of the AJT were rather weak for a research instrument, with .56 for plausible arguments and .54 for implausible arguments. For Part 1, internal consistency was at least acceptable (Cronbach's $\alpha = .64$). Furthermore, the stability of the AJT scores was tested with a small sample of psychology students ($N = 22$). Within a 13-month interval, AJT scores were highly correlated ($r = .60$). Given that students' competencies to evaluate informal arguments may be expected to increase with different gradients over the course of a year of studying, this correlation is quite substantial and provides evidence of predictive validity.

Rasch analyses

We also examined whether the items in the AJT conform to a Rasch model (Rasch, 1960). A one-parameter logistic model (1-PL model) for response accuracy in the AJT as dependent variable was estimated using the TAM (Kiefer, Robitzsch, & Wu, 2016) and the ltm (Rizopoulos, 2018) packages for *R* (R Core Team, 2016). Weighted likelihood estimates (WLE) were used for estimating person abilities (Warm, 1989). Rasch analyses are presented separately for AJT Part 1 (measure of the students' competence to accurately judge the plausibility of informal arguments), AJT Part 2 (measure of the students' competence to accurately assign argumentation fallacies), and for the combined AJT (measure of the students' competencies to evaluate the internal consistency of informal arguments). Item parameters for each model are shown in Table 1.

AJT Part 1

According to the Andersen likelihood ratio test that compares two partial samples divided at the arithmetic mean of the test values, the overall fit of the items to the 1-PL model was good, $\chi^2(df = 29, N = 225) = 33.34$, $p = .223$. The WLE reliability coefficient was .49, indicating a quite low reliability. However, weighted mean-square values as measures for the fit between predicted and observed responses ranged from 0.96 to 1.04 ($M = .99$, $SD = 0.02$), indicating a reasonable item fit ($t < 0.75$). Figure 2 shows the item–person map, in which person abilities (i.e., log-odds for the correct responses across items) and empirical item difficulties (i.e., log-odds ratios for correct responses across participants) are placed on a logit-transformed scale of response accuracy. For the first part of the AJT, the majority of the items were located in the lower range of the distribution, implying that they were relatively easy.

AJT Part 2

For Part 2 of the AJT, the overall fit of the items to the 1-PL model was not good, $\chi^2(df = 9, N = 225) = 20.40$, $p < .05$. Moreover, WLE reliability was low with a score of .53. However, weighted mean-square values again revealed an acceptable fit between predicted and observed responses with a range from 0.89 to 1.10 ($M = .99$, $SD = 0.06$) and $t < -1.74$. The item–person map for the AJT Part 2 (Figure 3) shows that the empirical item difficulties were mainly located in the upper range of the distribution. This indicates that the items in the second part of the AJT were relatively difficult to solve.

Table 1. Item parameters and standard errors for the Rasch analyses of the AJT Part 1, the AJT Part 2, and the combined AJT

		AJT Part 1		AJT Part 2		Combined AJT	
Item	Item type	Item difficulty	*SE*	Item difficulty	*SE*	Item difficulty	*SE*
1	Plausible	−3.40	.36	–	–	−3.50	.36
2	Implausible	−0.86	.15	0.33	.15	0.30	.14
3	Plausible	−0.58	.14	–	–	−0.61	.14
4	Plausible	−0.90	.15	–	–	−0.94	.15
5	Plausible	−1.76	.19	–	–	−1.83	.19
6[a]	Plausible	–	–	–	–	–	–
7	Implausible	0.21	.14	1.19	.16	1.08	.16
8	Plausible	−2.07	.21	–	–	−2.15	.21
9	Implausible	−0.65	.14	0.67	.15	_[b]	_[b]
10	Plausible	−1.70	.18	–	–	−1.76	.19
11	Implausible	−2.26	.22	−0.52	.15	−0.46	.14
12	Plausible	−0.73	.14	–	–	−0.76	.15
13	Plausible	−0.86	.15	–	–	−0.89	.15
14	Plausible	−2.59	.25	–	–	−2.68	.26
15	Implausible	−2.03	.20	−0.01	.15	−0.01	.14
16	Plausible	−0.94	.15	–	–	−0.98	.15
17	Plausible	−1.95	.20	–	–	−2.02	.20
18	Plausible	−0.97	.15	–	–	−1.01	.15
19	Plausible	−1.42	.17	–	–	_[b]	_[b]
20	Implausible	1.16	.16	2.22	.21	2.02	.20
21	Plausible	−2.47	.24	–	–	−2.56	.24
22	Plausible	−2.21	.22	–	–	−2.29	.22
23	Implausible	−0.20	.14	1.24	.17	1.13	.16
24	Plausible	−1.73	.18	–	–	−1.79	.19
25	Plausible	−1.51	.17	–	–	−1.56	.18
26	Implausible	−0.69	.14	0.07	.15	_[b]	_[b]
27	Plausible	−1.84	.19	–	–	−1.90	.19
28	Implausible	−0.58	.14	0.16	.15	0.15	.14
29	Plausible	−2.03	.20	–	–	−2.10	.21
30	Implausible	0.75	.15	3.53	.33	3.26	.33

Note. AJT Part 1 = plausibility judgements for 20 plausible and 10 implausible informal arguments. AJT Part 2 = recognition of typical argumentation fallacies for the 10 implausible informal arguments. Combined AJT = plausibility judgements for 20 plausible informal arguments and plausibility judgements as well as recognition of typicalargumentation fallacies for the 10 implausible informal arguments.

[a]Item was excluded from data analyses because it could have been wrongly interpreted in the sense of an argumentation fallacy.

[b]Items were excluded from the Rasch analysis because they appeared to be problematic after graphic inspection and application of Wald's test for differential item functioning.

Combined AJT

The two parts of the AJT measure the students' competencies to detect implausible arguments (Part 1) and to identify specific argumentation fallacies (Part 2). The combined AJT measures the students' general competence to evaluate informal arguments. For this analysis, responses to plausible arguments were scored as correct when students accurately judged them as plausible. Responses to implausible items were scored as correct only if students accurately judged them as implausible and also recognized the

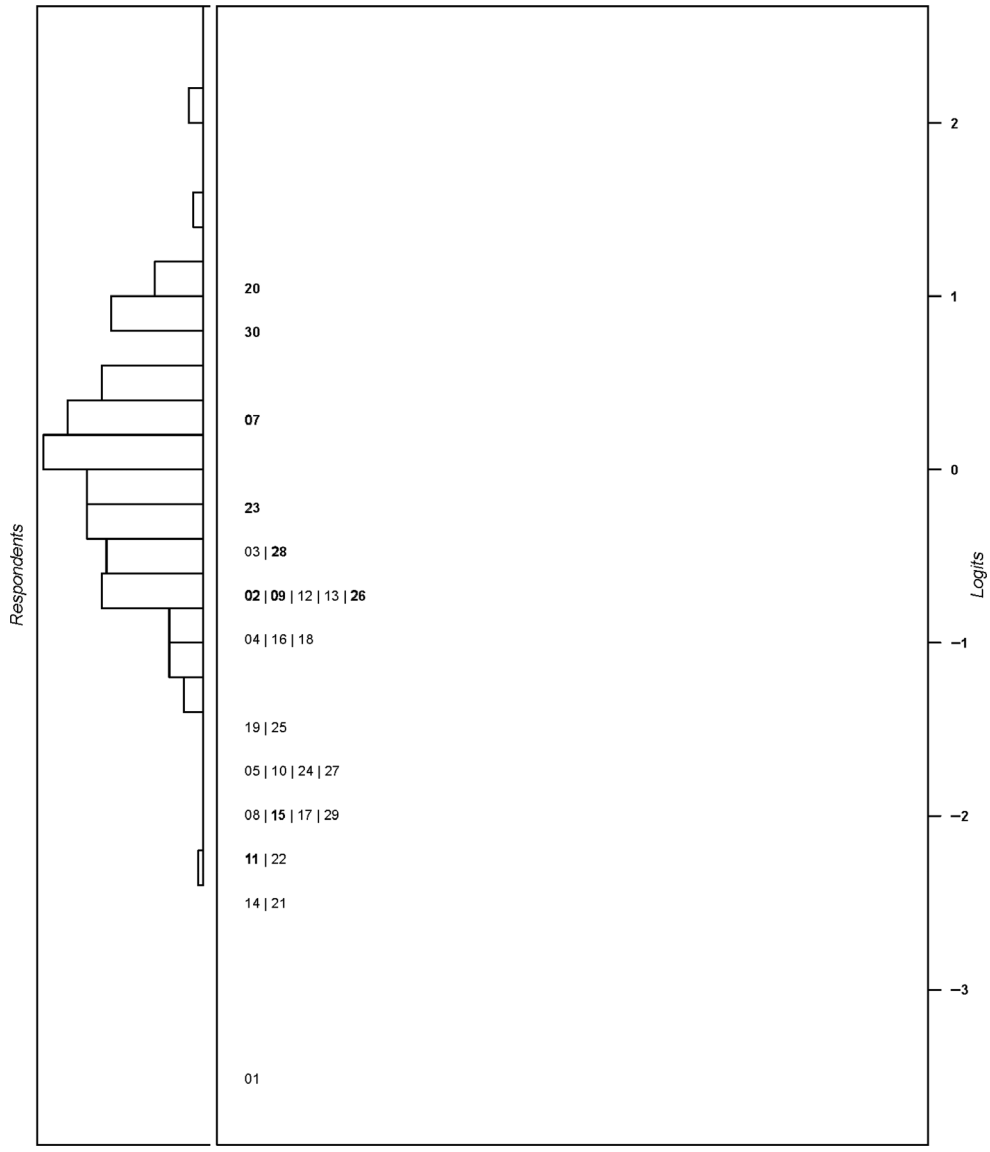

Figure 2. Item–person map for logit-transformed response accuracy in Part I of the AJT. Bolded items represent implausible arguments. The other items are plausible arguments.

corresponding argumentation fallacy correctly. After graphic inspection and applying Wald's test for differential item functioning, three test items had to be excluded from the analyses because they appeared to be problematic (Table 1). With the remaining items, the Rasch analysis revealed a good overall fit of the items to the 1-PL model, $\chi^2(df = 25, N = 225) = 27.53, p = .330$. The WLE reliability coefficient for this model was acceptable (.63) and weighted mean-square values were between .93 and 1.05, indicating a good item fit ($M = .99$, $SD = 0.03$, $t < -1.13$). For the combined AJT, average empirical item difficulties were only slightly below average person abilities, with the implausible items of the AJT being more difficult (Figure 4).

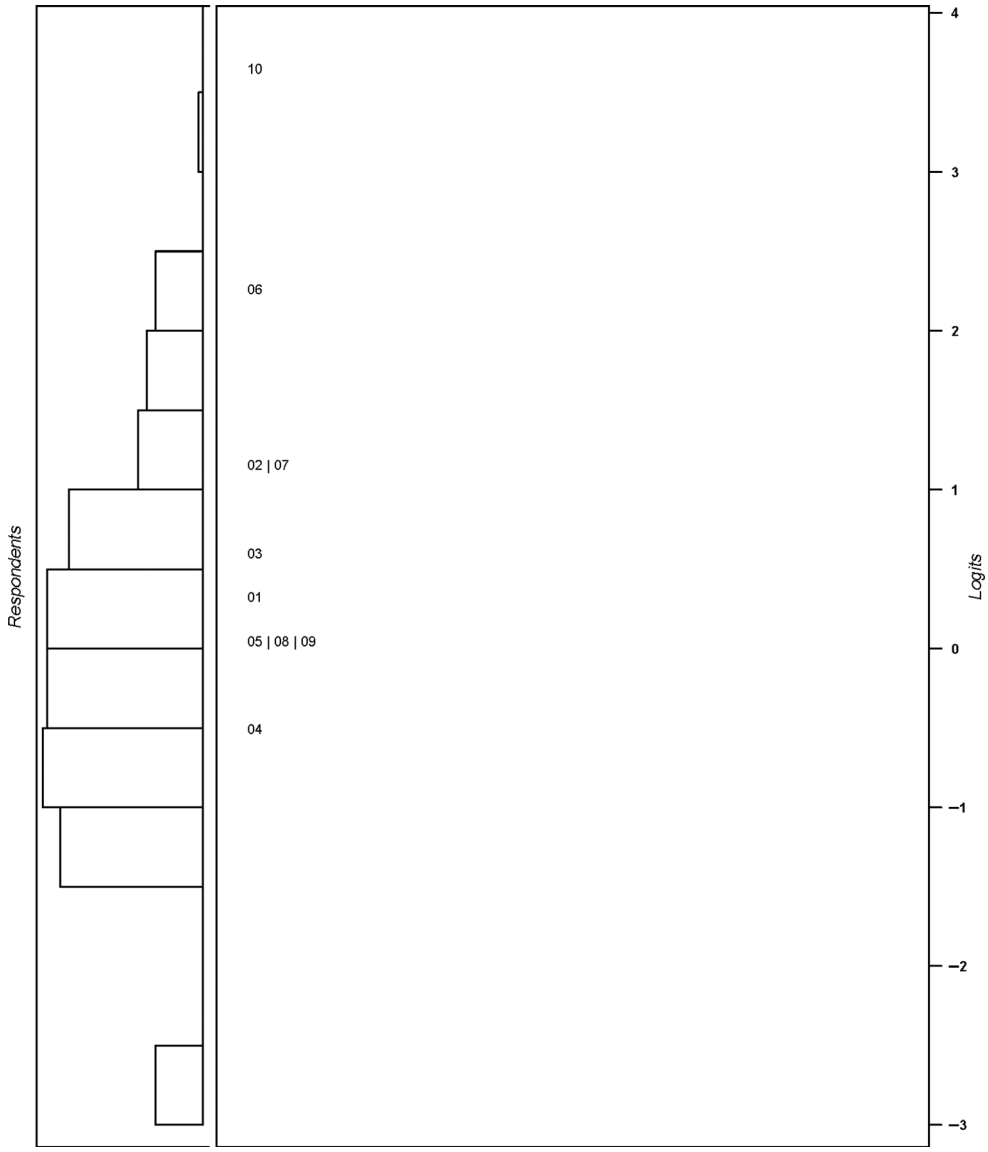

Figure 3. Item–person map for logit-transformed response accuracy in Part 2 of the AJT.

Effects of item properties on accuracy and response times

We further tested whether item difficulties depended systematically on item properties, which, theoretically, should facilitate or hamper the cognitive processes involved in argument evaluation. In particular, the distinction between plausible and implausible arguments was expected to be crucial in this respect. To clarify whether the evaluation of arguments, as targeted with the AJT, is a reflective activity that requires cognitive effort, we also estimated separate linear mixed models for response accuracy and response times as dependent variables and theoretically relevant item properties as predictors for Part 1 of the AJT.

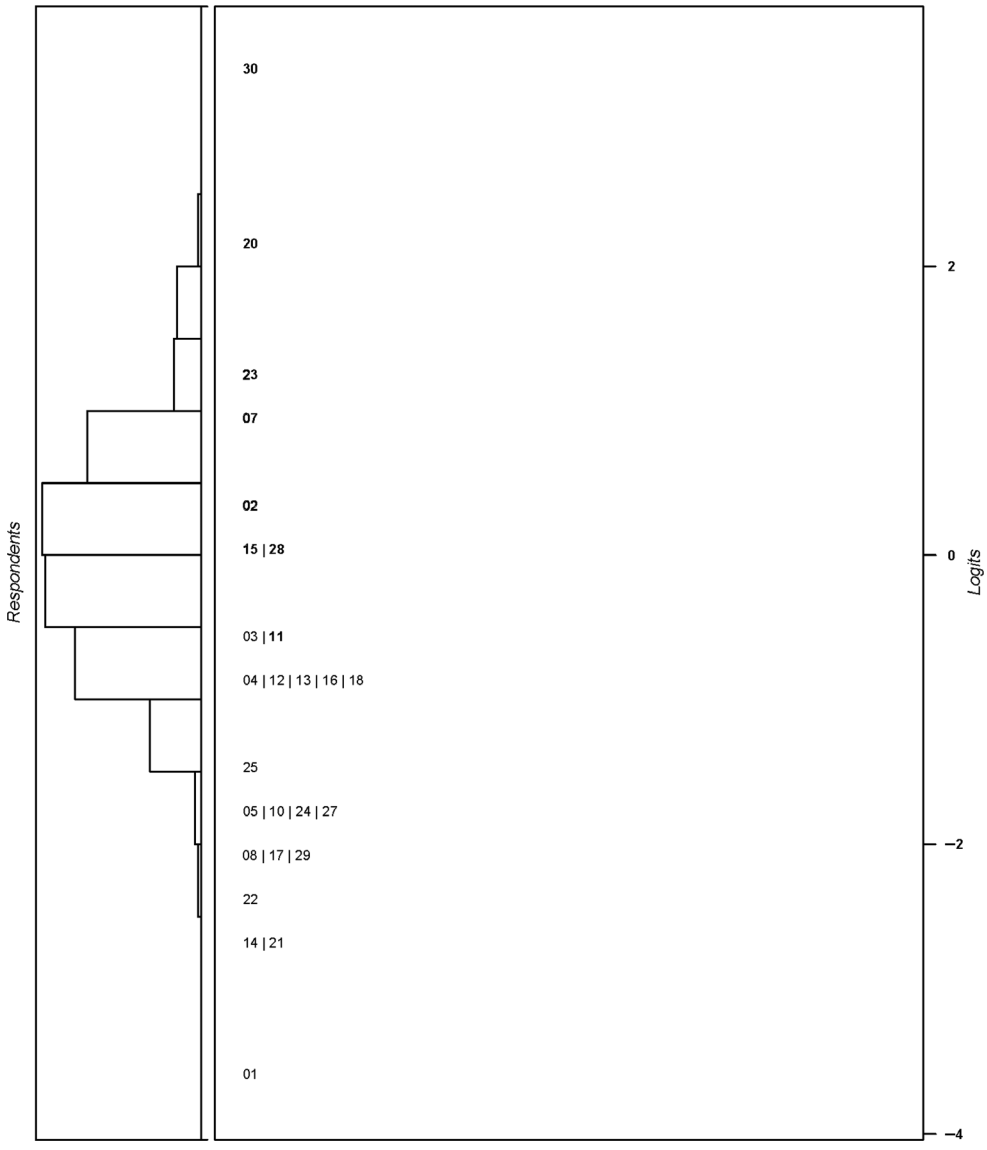

Figure 4. Item–person map for logit-transformed response accuracy of the combined AJT. Bolded items represent implausible arguments. The other items are plausible arguments.

Response accuracy

A generalized linear mixed model (logit link) with response accuracy in the first part of the AJT as dependent variable was estimated with the lme4 (Bates *et al.*, 2017) and lsmeans (Lenth, 2016) packages for R. Plausible arguments should be easier to evaluate than implausible arguments. Argument plausibility was contrast-coded (implausible argument = −1; plausible argument = 1) and entered as a fixed effect in the model. As control variables, argument length (number of characters without spaces for each argument, centred), information density of the arguments (number of content words for each argument, centred), and their interaction terms with argument plausibility were entered

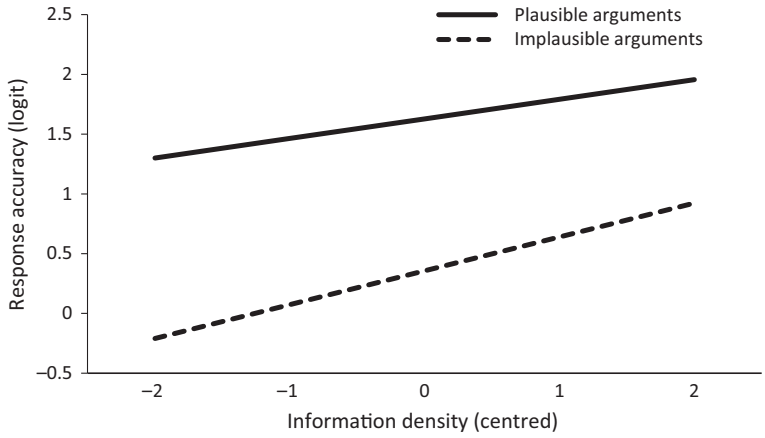

Figure 5. Response accuracy (proportion of correct responses in logits) in Part I of the AJT for plausible versus implausible arguments and information density (number of content words) of the arguments.

in the model (fixed effects). Intercepts and slopes were allowed to vary randomly between participants (random effects of participants) to account for the hierarchical structure of the data (items nested within participants). In line with Hypothesis 1, there was a main effect of argument plausibility (b = 1.271, SE_b = .092, z = 13.71, p < .001), indicating that plausible arguments were more likely to be judged accurately than implausible arguments. There were also main effects of argument length (b = −.022, SE_b = .002, z = −14.33, p < .001) and information density (b = .283, SE_b = .022, z = 13.12, p < .001). Thus, shorter arguments were easier to judge but, at the same time, arguments that contained more content words, that is, more information (when length was controlled for), were easier to judge as well. The model also revealed an interaction effect for argument plausibility and information density (b = −.120, SE_b = .033, z = −3.58, p < .001), indicating that the positive effect of information density was more pronounced in implausible arguments (Figure 5). There was also a significant interaction effect for argument plausibility and argument length in the reverse direction, b = .009, SE_b = .003, z = 3.53, p < .001.

Response times

A linear mixed model was estimated with response times in the first part of the AJT as dependent variable and argument plausibility (contrast-coded), response accuracy (effect-coded), argument length (centred), information density (centred), and the interaction of argument plausibility and response accuracy as predictors. Again, intercepts and slopes were allowed to vary randomly between participants (random effects). Considering that the detection and evaluation of implausible arguments are a rational and reflective process that requires cognitive effort, response times should be longer for implausible compared to plausible arguments if these arguments were correctly judged as implausible (Hypothesis 2). Again, main effects for semantic complexity (b = −193.74, SE_b = 48.73, t(6094.59) = −3.98, p < .001), argument length (b = 62.48, SE_b = 3.70, t(6098.33) = 16.88, p < .001), and argument plausibility (b = 1,689.25, SE_b = 366.79, t(1575.40) = 4.61, p < .001) emerged. There was also a main effect for response accuracy, indicating longer response times when the judgements were accurate, b = 2,152.06, SE_b = 351.8, t(5935.11) = 6.12, p < .001. However, the model also

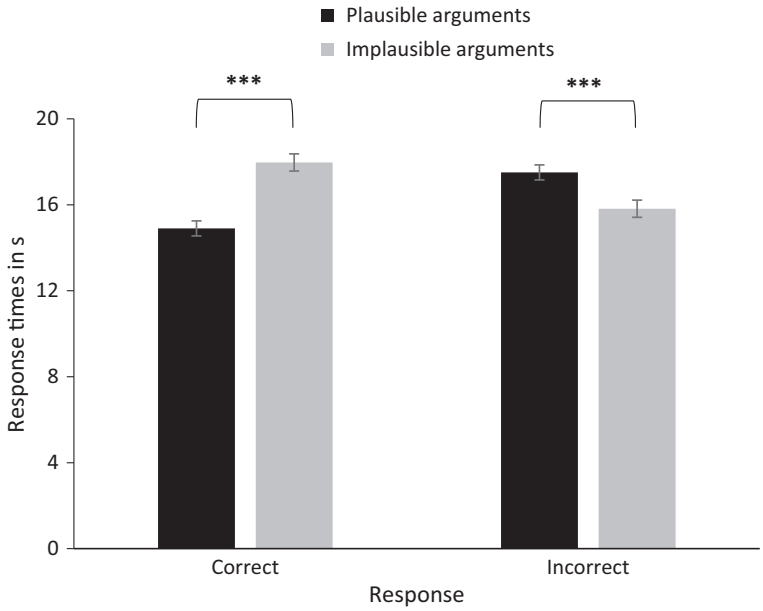

Figure 6. Response times (in seconds) in Part I of the AJT for plausible versus implausible arguments and response accuracy (proportion of correct responses). ***$p < .001$ (two-tailed).

revealed an interaction effect for response accuracy and argument plausibility, $b = -4{,}757.62$, $SE_b = 452.00$, $t(6211.68) = -10.53$, $p < .001$ (Figure 6). In line with Hypothesis 2, response times for implausible arguments were longer when the argument was judged correctly, whereas the opposite effect occurred for plausible arguments.

Correlations of argument evaluation with other constructs

In order to examine the nomological network of the ability to evaluate arguments in scientific texts, we inspected bivariate correlations of the test scores of the AJT with theoretically related constructs. We expected the AJT scores to be positively related to each other as well as to response times in the first task of the AJT, with argument comprehension measured with the AST, with the epistemological beliefs that psychological knowledge is structured but can nevertheless change, and with verbal intelligence (Hypothesis 3–7). Substantial correlations were found between the test scores of the AJT, as well as between the AJT and the AST and the variability subscale of the CAEB (lower values indicate stronger belief in variability) and verbal intelligence (Table 2). In sum, the pattern of the correlations of the AJT with theoretically related constructs indicates construct validity.

Relevance of argument evaluation for academic success

We assumed that the ability to evaluate arguments in scientific texts would be relevant for academic success at the university, implying a substantial correlation of the AJT scores with the current average grade of participants (Hypothesis 8). This prediction was supported (Table 2; note that the sign of the correlation is negative because lower grades indicate better performance in the German grading system).

Table 2. Correlations of the test scores in the argument judgment test (AJT) with other performance scores and epistemological beliefs

		M	*SD*	1	2	3	4	5	6	7	8[a]	9[b]	10
1	AJT 1	.74	.10	1									
2	AJT 1 *d'*	1.11	.66	.954***	1								
3	AJT 2	.36	.21	.662***	.730***	1							
4	Combined AJT	.66	.12	.880***	.799***	.696***	1						
5	AJT 1 RT	16.76	4.96	.199**	.230**	.241***	.180**	1					
6	AST	.67	.16	.339***	.349***	.427***	.374***	.176**	1				
7	Verbal IQ	.52	.12	.342***	.382***	.516***	.361***	.046	.403***	1			
8	GPA[a]	2.11	1.29	−.019	−.029	−.202**	−.106	−.070	−.166*	−.231***	1		
9	Current grade average[b]	2.02	.50	−.324**	−.314**	−.456***	−.412***	.159	−.198	−.349**	.251*	1	
10	CAEB-T	5.34	.99	.185**	.204**	.126	.148	.008	.204**	.155*	−.002	−.186	1
11	CAEB-V	2.71	.77	−.349***	−.359**	−.351***	−.325***	−.030	−.325***	−.197**	.086	.240*	−.474***

Note. AJT 1 = performance in Part 1 of the AJT (proportion of correct responses). AJT 1 *d'* = performance in Part 1 of the AJT (*d'* statistics). AJT 2 = performance in Part 2 of the AJT (proportion of correct responses). Combined AJT = performance in the combined AJT (proportion of correct responses in AJT Part 1 and AJT Part 2). AJT 1 RT = average response latency per argument in Part 1 of the AJT (in seconds). AST = performance in the argument structure test (proportion of correct responses, Münchow *et al.*, in press). GPA: grade point average (average grade in high school leaving certificate). CAEB-T/V = mean scores in the texture/variability subscales of the questionnaire Connotative Aspects of Epistemological Beliefs (Stahl & Bromme, 2007).

[a]$N = 224$.

[b]$n = 77$.

*$p < .05$; **$p < .01$; ***$p < .001$ (two-tailed).

Table 3. Hierarchical multiple regression analyses predicting academic success with measures of verbal intelligence, school leaving grade, and AJT performance scores

Predictor	Model 1 β	Model 2a$_1$ β	Model 2a$_2$ β	Model 2b β	Model 3a β	Model 3b β
Verbal IQ	−0.309^{+}	−0.199^{+}	−0.203^{+}	−0.84	−0.083	−0.084
GPA	0.183^{+}	0.207	0.201^{+}	0.149	0.157	0.143
AJT 1		−0.236^{+}			0.045	
AJT 1 d'			−0.212^{+}			0.040
AJT 2				−0.379^{++}	−0.337^{+}	−0.402^{+}
R^2	.15	.20	.19	.20	.24	.24
F	6.71**	6.00**	5.63**	7.49***	5.57***	5.56**

Note. n = 77. Standardized regression coefficients. GPA: grade point average (average grade in high school leaving certificate). AJT 1 = performance in Part 1 of the AJT (proportion of correct responses). AJT 1 (d') = performance in Part 1 of the AJT (d' statistics). AJT 2 = performance in Part 2 of the AJT (proportion of correct responses).
*p < .05; **p < .01; ***p < .001 (two-tailed).
^{+}p < .05; ^{++}p < .01 (one-tailed).

We also estimated a series of hierarchical regression models to test whether the ability to evaluate arguments explains variance in academic success over and above verbal intelligence and GPA (Hypothesis 9). In Model 1, verbal intelligence and GPA entered simultaneously had significant effects on academic success (Table 3). Importantly, however, the proportion of the correct responses (Model 2a$_1$) as well as the d'-test score (Model 2a$_2$) for Part 1 of the AJT (detection of implausible arguments) and the test score for Part 2 (identification of the correct type of fallacy, Model 2b) entered in the second step had significant increments in explaining academic success. The increment of the test score for the identification of the correct type of fallacy was larger (ΔR^2 = .20). Moreover, the ability to identify the correct type of fallacy was the most powerful predictor and the only predictor that remained significant in the full Models 3a and b that included AJT Part 1 and AJT Part 2 scores, verbal intelligence, and GPA.

Discussion

In this study, university students' ability to evaluate arguments in scientific texts was assessed with a novel computer-based diagnostic instrument (AJT) and related to other constructs. The item–person maps of Part 1 of the AJT revealed that most of the items in the lower part of the difficulty distribution refer to implausible arguments, which, in line with our assumptions, indicates that implausible items were more difficult than plausible items. Moreover, it was found that response times were longer for implausible arguments, but only if they were recognized as implausible. Additionally, implausible arguments whose information density was higher were easier to judge. These findings support the idea that the ability to evaluate arguments, as measured with the AJT, rests on cognitively effortful potential processes.

Analyses of psychometric properties showed relatively low internal consistencies of both AJT subscales (but an acceptable internal consistency for the combined scale). However, it should be noted that reliability estimates based on internal consistencies represent a lower-bound estimate of the true reliability of a measure. Indeed, the AJT scores show a quite remarkable stability over a period of more than

1 year. We also found substantial correlations between the students' ability to judge the plausibility of informal arguments and a measure of argument comprehension that required participants to identify functional argument components. This relationship and the substantial correlations with verbal intelligence and students' epistemological beliefs establish a nomological net of the ability to evaluate arguments in scientific texts: argument evaluation is promoted by but not identical with verbal intelligence. It may be construed as an aspect of rationality (Stanovich, 2012; Stanovich & West, 1998) as well as scientific literacy (Britt *et al.*, 2014) that requires mature and functional epistemological beliefs. In particular, the belief that psychological knowledge is structured but also changing might serve as a kind of metacognitive insight that promotes the ability to evaluate informal arguments.

Interestingly, the AJT scores were not correlated with the students' GPA in the school leaving certificate received from the academic track high school (Gymnasium). As a null result, this finding should not be overinterpreted. Nevertheless, it might hint at the fact that the evaluation of informal arguments is not part of the gymnasium curriculum but must be acquired at the university as part of a more general disciplinary expertise (von der Mühlen *et al.*, 2016).

The present study also found that the students' abilities to recognize argumentation fallacies and to accurately judge the plausibility of informal arguments were predictive for academic achievement in terms of the students' current grades average, even if verbal intelligence and the students' GPA were controlled for. This finding lends further support to the interpretation that the ability to evaluate arguments is distinct from general cognitive ability. It also underlines the practical significance of this ability for studying at the university and suggests the need to explicitly train university students in argument evaluation. The readers should note, though, that only about one-third of participants provided their average grade and systematic biases due to the large proportion of missing values cannot be ruled out.

Despite the consistent and interpretable results of this study, three other limitations must be noted. First, the AJT is based on text materials from psychology and the sample of participants consists of students from psychology and other social and educational sciences. We do not know whether and to what extent the results generalize to other populations of students. Presumably, for students of other subjects, a comparable test based on different texts taken from their field of study must be developed to answer this question. Second, we were not able to draw a random sample of participants from the population of students of social and education sciences but based our study on a convenience sample of students from just two universities. Future studies should attempt to cross-validate the present results based on a sample that comes closer to a random sample of university students Third, it would be desirable to look at the relationships of the AJT with additional criteria and constructs, such as tests scores of the Argument Evaluation Test (Stanovich & West, 1997), to establish a clearer picture of the nomological net of the ability to evaluate arguments. The results of the present study suggest that the efforts required for this research might be well spent.

Acknowledgements

This research was supported by the German Federal Ministry of Education and Research (Grant no. 01PK15009B).

References

Adey, P., & Shayer, M. (1993). An exploration of long-term far-transfer effects following an extended intervention program in the high school science curriculum. *Cognition and Instruction*, *11*, 1–29. https://doi.org/10.1207/s1532690xci1101_1

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *I-S-T 2000 R – Intelligenz-Struktur-Test 2000 R [Intelligence Structure Test 2000 R]*. Göttingen, Germany: Hogrefe.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Sigmann, H. (2017). *lme4: Linear mixed-effects models using Eigen and S4*(R-package version 1.1–14) [Computer software]. Retrieved from: http://cran.r-project.org/package=lme4

Bierhoff, H. W. (1993). *Sozialpsychologie: Ein Lehrbuch* (3rd ed.) [Social psychology: A textbook]. Stuttgart, Germany: Kohlhammer.

Brehm, S. S., & Kassin, S. M. (1996). *Social psychology* (3rd ed.). Boston, MA: Houghton Mifflin and Company.

Britt, M. A., Richter, T., & Rouet, J.-F. (2014). Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist*, *49*, 104–122. https://doi.org/10.1080/00461520.2014.916217

Dauer, F. W. (1989). *Critical thinking: An introduction to reasoning*. New York, NY: Oxford University Press.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Fuchs, R., & Schwarzer, R. (1997). Tabakkonsum: Erklärungsmodelle und Interventionsansätze [Tobacco consumption: Explanations and interventions]. In R. Schwarzer (Ed.), *Gesundheitspsychologie: Ein Lehrbuch* (pp. 209–244). Göttingen, Germany: Hogrefe.

Green, D. W. (1994). Induction: Representation, strategy and argument. *International Studies in the Philosophy of Science*, *8*, 45–50. https://doi.org/10.1080/02698599408573479

Herkner, W. (1991). *Lehrbuch Sozialpsychologie* (5th ed.) [Textbook social psychology]. Bern, Switzerland: Huber.

Hofer, B. K., & Pintrich, P. R. (Eds.) (2002). *Personal epistemology: The psychology of beliefs about knowledge and knowing*. Mahwah, NJ: Erlbaum.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge, MA: Harvard University Press.

Kiefer, T., Robitzsch, A., & Wu, M. (2016). TAM: Test analysis modules. R package version 1.995-0. Retrieved from: https://cran.r-project.org/web/packages/TAM/index.html

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*, 1–33. https://doi.org/10.18637/jss.v069.i01

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Hove, UK: Psychology Press. https://doi.org/10.4324/9781410611147

Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal*, *30*, 359–377. https://doi.org/10.1080/0141192041000168968 9

Münchow, H., Richter, T., der, von Mühlen, S., Schmid, S., Bruns, K., & Berthold, K. (in press). Comprehension of arguments in scientific texts: Reliability and validity of the Argument Structure Test (AST). *Diagnostica*.

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, *328*, 463–466. https://doi.org/10.1126/science.1183944

R Core Team. (2016). *R: A language and environment for statistical computing*. Retrieved from: https://www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Richter, T., & Schmid, S. (2010). Epistemological beliefs and epistemic strategies in self-regulated learning. *Metacognition and Learning*, *5*, 47–65. https://doi.org/10.1007/s11409-009-9038-4

Rizopoulos, D. (2018). *ltm: Latent Trait Models under IRT. R package version 1.1-1*. Retrieved from: https://github.com/drizopoulos/ltm

Rosman, T., Peter, J., Mayer, A. K., & Krampen, G. (2018). Conceptions of scientific knowledge influence learning of academic skills: Epistemic beliefs and the efficacy of information literacy instruction. *Studies in Higher Education*, *43*, 96–113. https://doi.org/10.1080/03075079.2016.1156666

Rouet, J. F., Favart, M., Britt, M. A., & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction*, *15*, 85–106. https://doi.org/10.1207/s1532690xci1501_3

Schroeder, S., Richter, T., & Hoever, I. (2008). Getting a picture that is both accurate and stable: Situation models and epistemic validation. *Journal of Memory and Language*, *59*, 237–259. https://doi.org/10.1016/j.jml.2008.05.001.

Shaw, F. W. (1996). The cognitive processes in informal reasoning. *Thinking and Reasoning*, *2*, 51–80. https://doi.org/10.1080/135467896394564

Stahl, E., & Bromme, R. (2007). The CAEB: An instrument for measuring connotative aspects of epistemological beliefs. *Learning and Instruction*, *17*, 773–785. https://doi.org/10.1016/j.learninstruc.2007.09.016

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 343–365). New York, NY: Oxford University Press.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*, 342–357. https://doi.org/10.1037/0022-0663.89.2.342

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*, 161–188. https://doi.org/10.1037/0096-3445.127.2.161

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing the development of rationality. In H. Markovits (Ed.), *The developmental psychology of reasoning and decision-making* (pp. 7–35). London, UK: Psychology Press. https://doi.org/10.4324/9781315856568

Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E. M., & Berthold, K. (2016). Judging the plausibility of argumentative statements in scientific texts: A student-scientist comparison. *Thinking and Reasoning*, *22*, 221–246. https://doi.org/10.1080/13546783.2015.1127289

Voss, J. F., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction*, *1*, 337–350. https://doi.org/10.1016/0959-4752(91)90013-X

Warm, T. A. (1989). Weighted likelihood estimation of ability in Item Response Theory. *Psychometrika*, *54*, 427–450. https://doi.org/10.1007/BF02294627

Received 31 January 2019; revised version received 15 May 2019

Appendix: Translations of the items of the AJT

Item	Original German version	English translation
1	Die Entwicklung zum Raucher und später ggf. auch zum Nicht-Mehr-Raucher basiert auf dem Zusammenwirken einer Vielzahl sozialer, psychologischer und biologischer Faktoren	The development of a smoker, and possibly later into a non-smoker, is due to the interaction of a variety of social, psychological, and biological factors
2	Eine zentrale Rolle scheint dabei das Konstrukt der ererbten Nikotinsensitivität zu spielen. Dieses Konstrukt bezieht sich auf die Tatsache,	The construct of inherited nicotine sensitivity seems to play a central role here. This construct refers to the fact that some people

Continued

Appendix. (*Continued*)

Item	Original German version	English translation
	dass manche Menschen sensitiver auf Nikotin reagieren, weil sie sensibler auf Nikotin ansprechen	react more strongly to nicotine because they are more sensitive to nicotine
3	Mit ihm soll erklärt werden, warum manche Menschen – obwohl sie schon eine relativ große Zahl von Zigaretten geraucht haben – nicht vom Nikotin abhängig werden, während andere eine hochgradige Nikotinabhängigkeit entwickeln	It aims at explaining why some people – although they already have smoked a relatively high number of cigarettes – do not become addicted to nicotine, whereas others develop a strong addiction to nicotine
4	Die Frage nach der genetischen Prädisponiertheit für das Rauchen bezieht sich aber nicht nur auf differentielle Unterschiede bei der Nikotinsensitivität, sondern darüber hinaus auch auf angeborene Persönlichkeitsunterschiede	The question of a genetic predisposition for smoking does not only refer to differential variation of nicotine sensitivity but also to congenital personality differences
5	Auf der Grundlage der Eysenckschen Drei-Faktoren-Theorie wurde insbesondere ein positiver Zusammenhang zwischen dem Rauchen und dem Persönlichkeitsmerkmal„ Extraversion" postuliert und auch in einer Vielzahl von Studien empirisch nachgewiesen (z.B. Lipkus *et al*., 1994). Gleiches gilt auch für den Zusammenhang zwischen Rauchen und dem von Zuckerman postulierten Persönlichkeitsmerkmal„Sensation Seeking", der Suche nach immer neuen Reizen und Stimulationen	Based on Eysenck's three dimensions of personality, a positive correlation between smoking and the personality trait 'extraversion' has been postulated and empirically verified by several studies (e.g., Lipkus *et al*., 1994). The same condition is true for the relationship between smoking and the personality trait of 'sensation seeking' postulated by Zuckerman, which is characterized by a constant search for new excitement and stimulation
6[a]	Während in den Aneignungsphasen des Rauchverhaltens vor allem biopsychologische Einflussgrößen eine wichtige Rolle spielen, so scheint im Stadium der Aufrechterhaltung das Rauchverhalten in erster Linie eine Funktion intrapersonaler Faktoren zu sein	Although biopsychological factors play an important role in the acquisition phases, the maintenance phase of smoking behaviour seems to be primarily a function of intrapersonal factors
7	Eine häufig zitierte Theorie zur Erklärung der Nikotinabhängigkeit ist die Nikotinregulationstheorie. In einem Experiment wurden Gewohnheitsraucher über mehrere Wochen hinweg mit Zigaretten versorgt, die entweder stark oder schwach nikotinhaltig waren, ohne dass dies für die Probanden erkennbar war (Schachter, 1980). Die Probanden rauchten im Durchschnitt 25% mehr von den leichten Zigaretten als von den starken. Aus diesem Ergebnis folgt, dass der Nikotingehalt einer Zigarette den wichtigsten Effekt auf das Rauchverhalten hat	A frequently cited theory to explain nicotine dependence is the nicotine regulation theory. In a blind experiment, habitual smokers were provided with cigarettes containing either a high or a low dose of nicotine for several weeks (Schachter, 1980). On average, test persons smoked 25% more of the weak compared to the strong cigarettes. It follows from this result that the nicotine content of a cigarette has the most important effect on smoking behaviour
8	Wenn Kinder bzw. Jugendliche anfangen, darüber nachzudenken, einmal eine Zigarette zu probieren, werden erstmals auf das Rauchen	When children or adolescents start to think about trying a cigarette for the first time, smoking-related impressions and

Continued

Appendix. (*Continued*)

Item	Original German version	English translation
	bezogene Vorstellungen und Erwartungen herausgebildet. Flay, d'Avernas, Best, Kersell und Ryan (1983) bezeichnen dieses Stadium als die Phase der Vorbereitung (preparation)	expectations are established. Flay, d'Avernas, Best, Kersell and Ryan (1983) termed this stage the 'phase of preparation'
9[b]	So führt z.B. die Erwartung, dass einem das Rauchen Stresssituationen erleichtert, eher dazu, dass man in solchen Situationen raucht, als die Erwartung, dass das Rauchen in Belastungssituationen hilfreich ist	The expectation that smoking relieves stressful situations, for example, is likely to lead to smoking behaviour in these situations than the expectation that smoking is helpful in stressful situations
10	Da eigene Erfahrungen mit Zigaretten noch nicht vorliegen, basiert die Bildung solcher Erwartungen hauptsächlich auf der Beobachtung des Modellverhaltens der relevanten Bezugspersonen	Since there is no personal experience with cigarettes, the forming of such expectations is mainly based on observations of the relevant caregivers' behaviours
11	Wird bei den anwesenden Kindern eher eine negative Vorstellung des Zigarettenrauchens bekräftigt, dann lässt der Vater z.B. nach dem Essen erkennen, wie herrlich ihm jetzt die Verdauungszigarette schmeckt	When a negative impression of smoking cigarettes is reinforced with children being present, the father, for example, realizes how wonderful the inhalation of his cigarette tastes
12	In der Längsschnittuntersuchung von Dinh, Sarason, Peterson und Onstad (1995) ist gezeigt worden, dass solche positiven Vorstellungen von den Wirkungen des Rauchens bei Fünftklässlern ein signifikanter Prädiktor für das Rauchverhalten vier Jahre später sind und dass diese positiven Vorstellungen bei Fünftklässlern das künftige Rauchverhalten stärker beeinflussen als entsprechende Vorstellungen bei Siebtklässlern	In the longitudinal study of Dinh, Sarason, Peterson and Onstad (1995), it was shown that such positive impressions of the effects of smoking on fifth graders are a significant predictor of smoking behaviour 4 years later and that these positive impressions have a stronger influence on future smoking behaviour in fifth graders than corresponding beliefs in seventh graders
13	Mit dem Rauchen der ersten Zigarette gelangen Jugendliche in eine Experimentierphase. Da etwa 80% bis 90% aller Jugendlichen wenigstens einmal eine Zigarette rauchen, trägt dieses Experimentieren nicht den Charakter eines abweichenden Verhaltens, sondern eher den einer normativen Entwicklungsaufgabe	Adolescents enter an experimental phase when they smoke their first cigarette. Given that 80% to 90% of all adolescents smoke a cigarette at least once, this experimenting is not deviant behaviour but rather a normative developmental task
14	Kritisch für die Herausbildung des gewohnheitsmäßigen Rauchens ist nicht der Umstand, dass eine erste Zigarette geraucht wird, sondern die Art und Weise, wie anschließend das Erlebnis dieser ersten Zigarette kognitiv und emotional verarbeitet wird	Crucial for the development of habitual smoking is not the fact that a first cigarette is smoked but the way in which the experience of this first cigarette is cognitively and affectively processed
15	Aus der Tatsache, dass am Ende der Jugendzeit der Anteil der gelegentlichen und regelmäßigen Raucher zusammengenommen auf etwa 50% absinkt (BZgA, 1994), lässt sich die Schlussfolgerung ziehen, dass das Interesse am	Based on the finding that the proportion of occasional and regular smokers decreases to around 50% at the end of adolescence (BZgA, 1994), it can be concluded that interest in

Continued

Appendix. (*Continued*)

Item	Original German version	English translation
	Rauchen im Laufe der Jugendzeit sogar noch zunimmt	smoking actually increases during adolescence
16	Das Selbstbild oder Selbstkonzept ist ein Forschungsgebiet der Sozialpsychologie, bei dem es um das im Langzeitgedächtnis gespeicherte Wissen eines Menschen über sich selbst geht	The self-image or self-concept is a field of research in social psychology that focuses on knowledge about oneself stored in long-term memory
17	Die Theorie der objektiven Selbstaufmerksamkeit von Duval und Wicklund (1972; Wicklund, 1975) beschäftigt sich damit, was passiert, wenn wir unsere Aufmerksamkeit auf unser Selbstbild richten	The theory of objective self-awareness by Duval and Wicklund (1972) explains what happens when we focus attention on our self-image
18	Allgemein wird mit ~"objektiver Selbstaufmerksamkeit~" dabei ein Zustand bezeichnet, bei dem die Aufmerksamkeit nach innen, auf die eigene Person gerichtet ist	In general, 'objective self-awareness' is a condition in which the attention is directed inwards, towards the self
19[b]	Der Zustand der objektiven Selbstaufmerksamkeit hat nach der Theorie von Duval und Wicklund (1972) bestimmte Auswirkungen	According to the theory of Duval and Wicklund (1972), the state of objective self-awareness can affect a person's behaviours and mental states
20	Eine Auswirkung wird darin gesehen, dass durch die Ausrichtung der Aufmerksamkeit auf die eigene Person Diskrepanzen zwischen dem Selbstideal (Anspruchsniveau in verschiedenen Bereichen) und dem realistischen Selbstbild stärker bewusst werden, weil dadurch diese Unterschiede deutlicher wahrgenommen werden	An effect of focusing attention on oneself is that discrepancies between the self-ideal (aspiration level in various areas) and the realistic self-image become more conscious because these differences are perceived more saliently
21	Dies kann sowohl positive als auch negative Selbstbewertungen zur Folge haben, je nachdem ob man z.B. einen überraschenden Erfolg erlebt (positive Selbstbewertung), oder ob man seinen Ansprüchen nicht gerecht geworden ist (negative Selbstbewertung)	Objective self-awareness can result in both positive and negative self-assessment, depending on whether a person has experienced, for example, a surprising success (positive self-assessment) or whether a person has not lived up to his or her own expectations (negative self-assessment)
22	Die Theorie postuliert, dass objektive Selbstaufmerksamkeit Diskrepanzen, sowohl im negativen als auch im positiven Sinne, zwischen Selbstideal und Wirklichkeit hervorhebt	The theory postulates that objective self-awareness highlights discrepancies, both negative and positive, between self-ideal and reality
23	Wenn eine positive Diskrepanz vorliegt, entstehen positive Emotionen, aber andererseits auch eine positive Selbstbewertung	If a positive discrepancy exists, positive emotions arise but also a positive self-assessment
24	Eine wichtige Hypothese der Theorie der objektiven Selbstaufmerksamkeit lautet, dass man im Zustand der objektiven Selbstaufmerksamkeit versucht, Diskrepanzen	An important hypothesis of the theory of objective self-awareness is that in the state of objective self-awareness, a person tries to reduce discrepancies between aspirations and

Continued

Appendix. (*Continued*)

Item	Original German version	English translation
	zwischen Anspruch und Wirklichkeit zu reduzieren, z.B. durch Anpassung des Verhaltens an die eigenen Einstellungen und Normen	reality, for example, by adapting behaviour to the person's attitudes and norms
25	Die Vorhersagen der Theorie der objektiven Selbstaufmerksamkeit zur Wahrnehmung von Diskrepanzen zwischen verschiedenen Aspekten des Selbst sind überwiegend in Form experimenteller Untersuchungen überprüft worden, indem die Versuchspersonen Reizen ausgesetzt werden, die die Aufmerksamkeit auf die eigene Person lenken, z.B. Spiegel	Based on the theory of objective self-awareness, studies investigating the perception of discrepancies between different aspects of the self have been tested mainly in experiments by exposing participants to stimuli that draws attention toward the self, for example, by using mirrors
26[b]	In einem Experiment von Ickes *et al.* (1973) bearbeiteten die Versuchspersonen zunächst verschiedene leistungsthematische Aufgaben. Im Anschluss gab ihnen der Versuchsleiter eine extrem positive Rückmeldung (z.B. einen deutlichen Tadel) über ihre Aufgabenbearbeitung	In an experiment by Ickes *et al.* (1973), the participants first worked on various performance-related tasks. Subsequently, the experimenter gave them an extremely positive feedback (e.g., a clear rebuke) about their task processing
27	Nach der Rückmeldung füllten die Versuchspersonen einen Selbsteinschätzungsfragebogen aus, indem eigene Leistungen und Fähigkeiten beurteilt werden sollten. Die Hälfte der Versuchspersonen saß dabei vor einem Spiegel, die andere Hälfte vor einer Wand	After feedback, the participants completed a self-assessment questionnaire to evaluate their own achievements and abilities. Half of the participants sat in front of a mirror, the other half in front of a wall
28	Im Ergebnis beurteilten sich die Versuchspersonen in der Spiegel-Bedingung positiver als die Versuchspersonen ohne Spiegel. Die Theorie von Duval und Wicklund (1972) wurde somit zweifelsfrei bewiesen	As a result, the participants assessed themselves more positively in the mirror condition than participants without mirrors. The theory of Duval and Wicklund (1972) was thus proven beyond doubt
29	Ein Erfolgserlebnis scheint den Selbstwert also dann zu steigern, wenn die gute eigene Leistung zum Selbstideal in Beziehung gesetzt wird	A sense of achievement seems to increase self-esteem when a person's good performance is related to the person's self-ideal
30	Aus den Ergebnissen der Untersuchung lässt sich außerdem schlussfolgern, dass, wenn die Versuchspersonen anstatt eines positiven Feedbacks ein negatives Feedback erhalten hätten, dies den eigenen Selbstwert geschwächt hätte	One conclusion from the results of the study is that if participants had received negative feedback instead of positive feedback, their self-esteem would have been weakened

Note. The items were originally presented in German language but have been translated into English in order to present the general content of the AJT. Please note that the English items have not been psychometrically tested.

[a]This item was excluded from data analyses because it could have been wrongly interpreted in the sense of an argumentation fallacy.

[b]These items were excluded from the Rasch analysis in the combined model because they were problematic after graphic inspection and application of Wald's test for differential item functioning.