Differentiating Theories from Evidence: The Development of Argument Evaluation Abilities in Adolescence and Early Adulthood

PETRA BARCHFELD
BEATE SODIAN

Ludwig-Maximilians-Universität München
Leopoldstrasse 13
80802 München
Germany

barchfeld@psy.lmu.de
*sodian@psy.lmu.*de

Abstract: An argument evaluation inventory distinguishing between different levels of theory-evidence differentiation was designed corresponding to the levels of argument observed in argument generation tasks. Five scenarios containing everyday theories about a social problem, and arguments to support those theories were presented to 170 participants from two age groups (15 and 22 years) and different educational tracks. Participants had to rate the validity of arguments proposed by a story figure, to support the theory, to choose the best argument, and to justify their choice.

The rating task proved to be very difficult for all age groups, with only 49% of the university students consistently rating valid evidence-based arguments higher than flawed arguments. Competence improved with age and educational level. In the choice task more than 80% of the adults preferred an argument that reflected theory-evidence differentiation over mere theory elaboration or flawed reasoning. However, only adults with a university education were able to also explicitly justify their choice. Overall, these findings imply that laypersons have similar conceptual problems in differentiating theory from evidence as it has been reported for evidence generation tasks (Kuhn, 1991). Performance on the choice task suggests that some implicit awareness of differences between theory and evidence may precede a full, explicit understanding. Implications for education are discussed.

Résumé: On a construit un questionnaire dans le but de mesurer différents niveaux d'habileté de distinguer des théories et des arguments. On présente cinq scénarios dans lesquels se trouvent des théories de tous les jours de problèmes sociaux et des arguments qui appuient ces théories à 170 participants de deux groupes d'âge (15 et 22 ans). Ceux-ci devaient évaluer la validité des arguments proposés par un personnage dans ces scénarios, appuyer la théorie, choisir le meilleur argument, et justifier leurs choix.

L'évaluation s'est avérée très difficile pour tous les groupes d'âge, car seulement 49% des universitaires identifiaient uniformément l'appui des arguments valides comme étant supérieur à l'appui des arguments défectueux. La compétence s'améliorait avec l'âge et le niveau d'éducation. Dans une des tâches plus de 80% des adultes préféraient plutôt un argument qui exprimait une différentiation théorie-appui que l'élaboration d'une simple théorie ou d'un

 Informal Logic, Vol. 29, No. 4 (2009), pp. 396-416.

argument incorrect. Toutefois, seulement des adultes avec une éducation universitaire pouvaient justifier explicitement leurs choix. Dans son ensemble, ces résultats impliquent que des non experts ont des problèmes conceptuels semblables à distinguer une théorie d'un appui, comme il l'a déjà été rapporté dans des études où les participants devaient construire des arguments (Kuhn, 1991). Leurs performances suggèrent qu'un conscience implicite entre une théorie et un appui puisse précéder une compréhension complète et explicite. On discute des implications éducatives.

Keywords: Argument evaluation, argumentation, development, theory evidence coordination, evidence evaluation

1. Introduction

Argument evaluation is a critical part of everyday reasoning. Typically, everyday reasoning is informal and inductive, and involves the evaluation of evidence to support a claim or conclusion within a given problem-solving context. The problem often is ill-defined and requires the use and evaluation of relevant empirical evidence (Means & Voss, 1996). While the ability to *generate* arguments to support or to refute a claim in informal reasoning contexts has been studied in some depth (e.g., Kuhn, 1991, Brem & Rips, 2000, Sá, Kelley, Ho & Stanovich, 2004), little work has been done on argument *evaluation*. From a developmental perspective, it appears important to ask whether argument evaluation abilities may precede argument generation in children and adolescents. As a first step towards a developmental investigation of argument evaluation, in the present study, an inventory was developed to study the ability to distinguish between different levels of evidence-based reasoning in adolescence and young adulthood.

Generating arguments

Skills of argumentation have been shown to be severely deficient in adolescence and even in adulthood (Kuhn, 1991; Kuhn, 2005; Kuhn & Franklin, 2006). In particular, people often fail to generate evidence relevant to evaluating a claim they are making. Rather, they offer an example or a script-like elaboration of their theory. Successful argument production requires reflective access to beliefs or theories a person holds with some conviction. Based on an in-depth investigation of informal argumentation in adolescents and adults, Kuhn (1991) argued that laypersons often fail to make their own theories an object of conscious evaluation and that one of the reasons for their failure to do so is an undifferentiated concept of theories and evidence.

In Kuhn's (1991) seminal study, participants were requested to develop a causal theory concerning a common social problem like

why some children fail at school or why some prisoners return to crime after being released from prison. This was followed by two sets of questions. The first consisted of three requests to generate evidence to support the subjective theory. After that, participants were asked for an alternative theory for criminal recidivism, for evidence supporting that alternative, and for evidence against the alternative, thus supporting their own theory.

Only a minority of the adult subjects in Kuhn's studies (from 9% to 22%) consistently showed a mature epistemological understanding, that is, the ability to systematically evaluate their own subjective theories, to consider alternatives to their own theory and to generate, contemplate and evaluate arguments for and against alternative theoretical positions. Most people believed that their own preferred theoretical explanations were true, in the naïve sense of never even having reflected on the possibility that there could be other explanations for the phenomena in question. Consequently, many participants who held this absolutist epistemological view did not see the point in critically evaluating their own or others' theories. Only 16% of Kuhn's subjects consistently generated genuine evidence for their own theories. Many participants simply elaborated their theories, and some (about 30%) generated what Kuhn called pseudoevidence, a descriptive instance or example that merely elaborates the theory that is taken to be true. Similarly, only about one third of Kuhn's subjects consistently generated alternative theories, and was consistently able to generate a counterargument either to their own or to an alternative theory. Adolescents and older adults performed worse than young adults, and performance covaried with educational level. This is consistent with a large body of findings indicating that education, rather than age, appears to be responsible for developmental differences in epistemological understanding (Kuhn, 1991; Kuhn, Weinstock & Flaton, 1994, Kuhn & Felton, 2000; Leadbeater & Kuhn, 1989).

In sum, Kuhn's research on the skills of argument revealed a metacognitive understanding of the theory-evidence relation in informal reasoning processes only in a minority of adolescents and adults. Subsequent research has generally yielded results consistent with these findings. Brem and Rips (2000) argued that the abilities of Kuhn's subjects may have been underestimated since they had no access to empirical data relevant to evaluating the complex social problems in question. When prompting subjects to imagine the strongest supporting evidence one could provide for a given theory, Brem and Rips found genuine evidence production in 68% of their participants. Still, adults performed far from ceiling. Similarly, Barchfeld (2008) found that young adults could be helped to think of genuine evidence to support their theory, but only 50% of a sample of N=151 22-year-olds developed arguments based on empirical evidence even after a series of very specific

prompts (Bullock, Sodian, & Koerber, 2009). A study by Sá et al. (2005) indicated that 74% of the adult participants were unable to create a covariation comparison in an argument.

In a study by Glassner, Weinstock and Neuman (2005), eighth-graders were presented with 6 scenarios including issues like whether or not there is life on other planets. The task was to generate two assertions for each problem. One was to prove the claim (e.g. "In your opinion what would be the best proof for the claim that there is life on other planets"), the second was to explain the phenomenon in question (e.g., "assuming that the earth is getting warmer, what would be the best explanation for that fact?"). Most of the participants correctly generated explanations when requested to explain a phenomenon, but the majority also generated explanations, rather than evidence, when asked to prove a claim. Thus, research conducted with different methods indicates that the cognitive decoupling of theory and evidence in argument production poses a great difficulty for adolescents and under some conditions even for educated adults.

Argument production is related to the degree of epistemological understanding: Studies by Weinstock and Cronin (2003) on juror reasoning found that epistemological understanding was a predictor of individual differences in the quality of argument production, independently of IQ. However, it can be argued that both argument production and epistemological reasoning involve high verbal demands. It is possible that laypersons fail to produce a valid argument, even though they can recognize one when given a judgment task. This may especially hold for children and adolescents, as well as for participants with limited verbal skills.

Judging Arguments

Little research has been reported on laypersons' ability to judge the validity of evidence-based arguments in informal reasoning contexts. Kuhn & Pearsall (2000) explored whether 4- to 6-year-old children could judge their source of knowledge when given a choice between a piece of evidence, and a cause for the outcome of an event. For example, pictures were given of a race with a cue for who won (boy holding the trophy) and a cue for the possible reason (shoe type). When asked who won the race, the children responded that the boy with the trophy won. When asked how they knew, however, they responded in an explanation based manner, referring to the shoes (why he won) rather than the trophy (how you know he won). The 6-year-old children distinguished between the two justifications more readily than 4-year-olds, who tended to merge the two justifications.

Another similar study (Kuhn & Felton, 2000) showed that while children as young as six years may be able to recognize a

piece of evidence, an explicit understanding of the validity of evidence and argument strength was poor even in university students. Eighth graders, college students and beginning graduate students, were asked to choose the stronger one of two arguments in support of a claim. One argument provided a theoretical explanation that made the claim plausible (Why is it so?), whereas the other provided empirical evidence that the claim was true (How do you know?). More important than the choices were the reasons participants gave for justifying their responses. They where asked if the chosen argument had any weaknesses and the nonchosen one had any strengths. The percentage of students citing the epistemic strengths of the explanation (e.g., "It gives a reason") ranged from 30% among the young teens to 60% among graduate students and the percentage of students citing the epistemic strengths of evidence (e.g., "It's something that really happened") ranged from 11% to 76% across groups. The performance in analyzing the weakness of explanation and evidence was much worse. Between 0% and 26% identified the epistemic weakness of explanation (e.g., "It's only a theory") and the fewest students (2% to 10%) that of evidence (e.g. "It doesn't say why").

The study by Glassner, Weinstock and Neuman (2005) mentioned above examined in a second task high-school students' ability to distinguish between explanation of a theory and development of evidence to support a theory with reference to the goal of the particular argument situation. Their participants, eighth-grade pupils, read 6 argumentation scenarios each having a stated goal of either explaining or proving a claim. Participants rated the degree to which each of two provided arguments (one a theoretical explanation, one a piece of evidence) helped to achieve the goal of the argument. In a second step, participants chose which one of the two arguments should be more effective in achieving the stated goal. The findings indicate a sensitivity to the relative epistemic strengths of explanation and evidence since participants rated explanations as more advantageous in achieving the explanation goal and evidence as more advantageous in achieving the proof goal.

In sum, it appears that children and adolescents are able to make a distinction between causal explanations for why something happened and evidence supporting a claim about what happened. However, their understanding of the epistemic strengths and weaknesses of the two types of justifications appears to be weak through high school, and even university students' explicit understanding of the distinction is deficient.

Aims of the Present Study

Although the studies reported above address a fundamental aspect of argument evaluation, they cannot be compared to argument generation tasks, since the types of valid, partially valid and invalid arguments generated spontaneously by the participants in Kuhn's (1991) study were not presented for evaluation. We do not know whether adults, who often produced theory elaborations rather than evidence in Kuhn's argumentation task, would *recognize* the difference between the two types of responses in a judgment task. Furthermore, the ability to distinguish between arguments varying in quality, such as genuine evidence versus pseudoevidence, has not been studied.

In the present study, we developed and tested an evidence evaluation inventory designed to assess the ability to distinguish between different levels of theory-evidence differentiation in children, adolescents, and adults. The model of levels of understanding the "Nature of Scientific Knowledge" derived from Carey et al. (1989) was used to develop the inventory. At the lowest level (Level 1) there is no differentiation between theory and evidence, as for example, when a justification is entirely based on theory elaboration or when facts are produced without any understanding of how these facts bear on a theoretical claim. At an intermediate level (2), there is partial differentiation between theory and evidence, for example, when single cases are mentioned to make a theory plausible or when empirical tests are suggested, but the quality of the evidence is poor and thus reflects an immature understanding of the notion of hypothesis testing. At Level 3, hypotheses are clearly differentiated from evidence, and valid tests are produced, but there is lack of an integration of single pieces of evidence into a coherent evaluation of a theory. At the highest level (Level 4), theory and evidence are differentiated and valid evidence is brought to bear on a theoretical claim. Typically, theory evaluation involves a series of empirical tests. A mature understanding of argumentation thus involves an awareness of alternative interpretations of individual pieces of evidence, and an understanding of the cumulative and cyclical nature of theory evaluation.

The evidence evaluation inventory was designed to reflect the main levels of responses observed by Barchfeld (Barchfeld and Sodian, in prep.) in an argument generation task modeled after Kuhn (1991). Thus, the task was designed to reflect a theoretically derived and empirically observed hierarchy of levels of epistemological understanding. Our main aim was to better characterize the nature of the deficit observed in many adolescents and adults in argument generation tasks. If the observed lack of theory-evidence generation is due to a deep-seated conceptual deficit, then similar difficulties should emerge in an evaluation task. If, on the other hand, the difficulties in spontaneous production are mainly

attributable to task demands such as knowledge (of relevant pieces of evidence) or verbal skills, then we should expect an evaluation task to pose no major difficulties to adult participants. Our second aim was to validate the theoretically postulated levels of theory-evidence differentiation, testing whether the empirically observed hierarchy of rated arguments would reflect the theoretically derived levels of difficulty. That is, we expected an age trend from adolescence to adulthood in the rating, choice and justification of arguments, and we expected evidence evaluation abilities to be related to educational level.

2. Method

Participants

The sample included 170 participants from two different age groups (ninth grade and early adulthood) and two different educational tracks (college-bound and vocationally oriented) from the German three-fold educational system. In this system, students are assigned to one of three educational tracks (Gymnasium, Realschule, Hauptschule) in the fifth grade. The lowest track ("Hauptschule") is vocationally oriented and ends in grade 9 or 10, at about the age of 15 to 16. A "Hauptschule" is a secondary school, starting after 4 years of elementary schooling. Any graduate of an elementary school can continue with Hauptschule, whereas a higher track secondary education requires good academic performance. The medium track ("Realschule") ends in grade 10. The Realschule is ranked between Hauptschule (lowest) and Gymnasium (highest). The Gymnasium is the university-bound educational track for gifted students. It ends with grade 12, and a high-school diploma (Abitur). The final two years at a Gymnasium are sometimes seen as comparable to the first two years in college in the United States.

In the present study, we compared students from the highest (Gymnasium) and lowest (Hauptschule) tracks. Our first group of 104 ninth graders (51 female, 53 male) with a mean age of 15.5 (*SD* = .74; Min. = 14.1 years, Max = 17.7 years) was recruited from two different school types. 56 (32 male, 24 female) attended "Hauptschule" (vocationally oriented) and 48 (21 male, 27 female) attended Gymnasium (college-bound).

The second group consisted of 66 participants (30 male, 36 female) with a mean age of 22.25 years (*SD* = .52; Min. = 21 years, Max = 23 years). They were recruited from a sample that had been studied longitudinally in the LOGIC study. This study (Longitudinal Study of the Genesis of Individual Competencies) was conducted by the Max-Planck-Institute for Psychological Research (Weinert & Schneider, 1999), assessing individual dif-

ferences in cognitive development (memory, thinking, arithmetic skills, mathematical understanding, reading and spelling, analytical reasoning in science) as well as personality development (motives, attitudes, academic self-concepts, social competencies, and moral judgment). The study started in 1984 when the participants were 3 to 4 years old and ended in 2005 when they were 22 years old. In the present study, 41 (20 male, 21 female) participants were university students at the time of testing, 22 (7 male 15 female) participants were employed or other (unemployed, at military service or other), and 3 (2 male, one female) had missing data regarding employment.

The Evidence Evaluation Inventory

The evidence evaluation inventory is a structured interview that consists of three sections. First, participants had to judge evidence of different validity supporting theories about six every day social problems. They were then requested to choose the most valid piece of evidence, and finally they were asked to justify their choice.

In support of each of the candidate causes four arguments of different quality were presented. The quality of arguments varied according to the scheme developed by Carey et al. 1989 (see Introduction). Level 0 was applied in the case of ignorance or irrelevant answers. At Level 1, theory and evidence are not differentiated, e.g., only theory elaboration is provided. At Level 2, observations and findings are brought to bear on a causal theory. However, it is unclear why and how these pieces of evidence are supposed to support the theory. At Level 3, there is a clear understanding of the notion of an empirical test for a claim relevant to theory evaluation, but individual tests are not embedded in an evaluation of the theoretical framework. At Level 4, there is a clear understanding that findings of an empirical test can be interpreted in different ways depending on theoretical viewpoints and the process of theory evaluation is seen as cyclic and cumulative. All arguments were formulated as a first person narrative.

In Table 1 the level system and an example for an insufficient argument (Level 1), a partially valid argument (Level 2) and a valid argument (Level 3 and Level 4) is reported. The arguments refer to the problem of increased smoking among adolescents and its candidate cause—passive smoking during childhood.

Participants were requested to rate each argument by applying school grades from 1 (very good evidence) to 6 (insufficient evidence; 2 = good evidence, 3 = satisfying, 4 = sufficient, 5 = faulty). In the German School system grades range from 1 for a very good performance to 6 for a insufficient performance. After rating each argument subjects were requested to choose the most valid argument: “What do you think, which person provided the

best proof?", and subsequently had to justify their choice in writing: "Why do you think so?" Participants had to justify each of their choices, but the justification was coded only if participants chose an argument on at least Level 3 (since arguments on Levels 1 and 2 can not be justified by referring to valid empirical evidence).

The responses to the justification question were coded into levels corresponding to the level of understanding described above. Level 0 was applied in the case of ignorance or irrelevant answers or in the case that none of the provided arguments was convincing the participant, e.g. subject 367 answers: "I think nobody produced good evidence." At Level 1 evidence was treated as a fact: The justifications just repeated the content of the evidence, referred to the competence of mentioned experts, referred to subjective plausibility, or just added some more examples or mechanisms to the chosen evidence (e.g., the justification of the math's evidence of participant 43: "Because person number three says that practice does not influence the performance"). Level 2 was applied when participants referred to the fact that there was empirical evidence without further elaborating why the evidence supported the theory (e.g., participant 67 for the smoking problem: "Because the person has an example, a boy, who started smoking at the age of 10"). At Level 3 subjects discussed the relevance of the method of generating evidence for the causal chain postulated by the theory (e.g., participant 592 wrote: "Because the person established a comparison and therefore can demonstrate that media can influence aggression and violence"). To perform at Level 4, participants had to elaborate the relevance of the method for testing the theory, including possible limitations of the evidence provided in an argument (e.g., subject 71 wrote: "The theory was tested with two groups (independent variables): movies with violent content for the one group and non violent movies for the other. The behavior after watching TV (dependent variable) functions as indicator for aggressiveness").

Procedure

The adolescents were tested in the classroom towards the end of the academic school-year. They read the scenarios by themselves and answered in writing. All participants were encouraged to ask if something in the interview was not clear to them. The administration time for the entire interview was about 30 minutes.

The sample of the young adults could not be interviewed personally therefore the questionnaire was sent to them by post with a prepaid envelope. Of the 151 contacted participants of the LOGIC Study 66 returned the completed questionnaire.

Table 1

Examples of Arguments for Each Level

Level	Prototypical Argument
Level 1: No differentiation of theory and evidence	
The argument consists of a further description of the theory by elaborating mechanisms consistent with the theory or by specifying new instances of the presumed causal chain. No genuine evidence is pro-vided for the theory, nor is there any reflection on possible ways of generating relevant evidence. Arguments at level 1 also included examples of two standard reasoning fallacies: The “ad populum” argument appeals to a systematic feeling of group solidarity in order to prove the point (“If most of the people believe this, then so do I”). In the “ad ignorantiam” argument a conclusion about the truth of a proposition is derived from the fact that the proposition is not known to be false.	That children were addicted to the smoke of their parents for many years must be a main reason anyway. Because there must be one reason and as long as there is no evidence for other reasons this must be a main reason. Therefore I am convinced that this is an important cause for the high consumption of cigar-ettes among adolescents.
Level 2: *Partial differentiation of theory and evidence*	
Observations and findings are brought to bear on a causal theory; however, it is unclear why and how these pieces of evidence are supposed to support the theory. For example, single cases of co-occurrence of cause and effect are cited, or observations that are not clearly related to the theory.	I know that because my friend is a social scientist. He told me of a boy whose parents were chain smokers. They used to smoke all day even at home. Since his birth the boy was surrounded by smoke. His first cigarette he smoked at the age of ten and since then he is smokes regularly. Meanwhile he smokes 2 packs a day. I think this is evidence for the claim that passive smoking is a main reason for the high consumption of cigarettes among adolescents.
Level 3: Simple differentiation of theory and evidence	

There is a clear understanding of the notion of an empirical test for a theory. However, evidence supporting simple beliefs, rather than more complex belief systems or theories is generated.	I know a study in which teenagers were compared according to their cigarette consumption. One group came from homes where parents were strong smokers and one from homes were parents did not smoke. The clear result was that children with parents who smoke were smokers themselves.
Level 4: Full differentiation of theory and evidence	
Arguments are based on an understanding of the target theory and its relation to alternative theoretical explanations. There is a clear under-standing that findings of an empirical test can be interpreted in different ways depending on theoretical view-points. Thus, a series of investi-gations is outlined to rule out alternative interpretations. For example, a set of empirical tests is laid out to test specific hypo-theses derived from the theory and to rule out alternative inter-pretations of a set of findings.	I know that because I read the results of a investigation in a report from the Ministry of Health: children who grew up in families in which both parents were smoking at least one pack of cigarettes a day, were smoking twice as much and did also start smoking earlier than children who grew up with parents with little or no cigarette consumption. Moreover the results indicate, that the influence of peers makes no difference in smoking behavior of children. I think that this good evidence for the importance of the parents' habits in early cigarette smoking.

3. Results

Rating of arguments

To examine whether participants are able to differentiate between arguments of different validity, the participants' ratings of the arguments were analyzed. Table 2 shows the mean rating of each level separately for each age and educational group. The data indicates for all age groups that they tended to rate the more valid arguments worse and the more invalid arguments better than the grade system would require. The appropriate grade for level 4 evidence (highest level of evidence) was defined as "1" (very good) or "2" (good). For level 3 evidence, "2" or "3" (satisfying) was assigned, for level 2 evidence "3" or "4" (sufficient), and for level 1 evidence "4", "5" (faulty) or "6" (inadequate). Especially the 15-year-olds showed little variation in their ratings (2.91min, 3.34 max). Furthermore, the 15-year-olds from the low track tended to rate invalid (level 1)

arguments better (2.94) than valid arguments (3.31). Only the university students seem to grasp the differences between levels (2.68 min, 4.59 max).

Table 2

Mean Ratings of different Level of Argument According to Age and Educational Level (Standard Deviations in Parentheses)

Cohort	Mean Rating			
	Level 1	Level 2	Level 3	Level 4+
15 yrs, Low Ed. Track	2.98 (.70)	2.94 (.76)	3.31 (.80)	3.22 (.85)
15 yrs, High Ed. Track	3.34 (.60)	3.05 (.56)	3.09 (.60)	2.91 (.69)
22 yrs, Occupied	3.78 (.90)	3.63 (.86)	3.20 (.86)	2.99 (.92)
22 yrs, Univ. Stud.	4.59 (.91)	4.30 (.93)	2.96 (.83)	2.68 (.82)

Table 3

Percentage of Correlations between Participants' Subjective Ranking of Arguments and the Optimum Ranking

	15 yrs, Low Track	15 yrs, High Track	22 yrs, Occupied	22 yrs, Univ. Stud.
No Correlation	100.00	90.00	75.00	27.50
Correlational Trend*	–	–	5.00	7.5
Sig. Correlation	–	10.00	5.00	22.5
Highly sig. Correlation	–	–	15.00	42.5

* Correlations between $r = .05$ and $r = .06$

In order to investigate if participants were able to rank the arguments appropriately (e.g., giving the best grade to a level 4 argument, and the worst grade to a level 1 argument and so on) the ranking resulting from the participants rating was compared with the optimal ranking of arguments for each participant (nonparametric correlation coefficient Kendall´s Tau). Table 3 shows the percentage of those response patterns that correlate with the optimal ranking of arguments, with the most valid argument being that with the highest grade and the least valid that with the lowest grade. The results indicate that if the subjective range of arguments is considered, none of the low track adolescents' and only 10% of the high track students' ranking showed a correlational pattern,

whereas 20% of the employed 22- year-olds and 65% of the university students put the arguments in the adequate order.

Choosing the Best Argument

Because many participants rated more than one argument of a problem with identical grades they were requested to choose one out of the four arguments for each problem as the most valid one. The percentage of participants who at least chose 5 out of 6 arguments correctly was 0% for the 15-year-old lower track adolescents, 3.6% for the 15-year-old higher track students, 13.4% for the 22-year-old employees, and 30.6% for the 22-year-old university students.

However, participants' understanding may have been underestimated by just taking correct choices of Level 4 arguments into account. Level 3 arguments also reflect the idea of empirical testing, and the distinction between Level 3 and 4 arguments is subtle. Moreover, not only the choice of Level 3 or 4 arguments, but also the rejection of Level 1 arguments can be taken as an indicator of a nascent understanding of the validity of evidence. We therefore calculated a less stringent competence score by defining a participant as being competent if he or she chose a level 4 or a level 3 argument on at least 50% of the tasks (i.e., at least 3 out of 6 for adolescents and adults), and if he or she did not choose a level 1 argument more than once. As a result, 9.1% of the low educational track 15-year-olds, and 36.5% of the high track 15-year-olds were scored as competent in their choice. In the 22-year-old group this was the case in 58.8% of the employed participants and 82.4% of the university students. Adults outperformed both groups of the 15-year-olds, X^2 (1, N=149) = 37.13, p< .00, Differences between educational levels were significant for the 15-year-olds, X^2 (1, N=96) = 9.90, p = .01, but only marginally significant for the adults X^2 (1, N=51) = 3.03, p = .07.

Justification of choice

To make sure that participants explicitly understand why they chose a particular argument, they had to justify their choice in writing. Performance on the justification task was coded according to whether or not participants referred to the evidence presented in the argument they chose (see above). Evidence-based justifications in response to at least three of six problems were provided by 68.4% of the university students. In the 22-year-old employed group, only 35.0% provided evidence-based justifications in at least three out of the six problems, X^2 (1, N=58) = 5.07, p < .05. In the 15-year olds, 19.2% of the higher track students and 15.9% of

the lower track students provided an evidence-based justification for at least half their choices, X^2 (1, N=99) = 0.33, p = .57. These findings imply that only in the group of university students the majority of participants showed a clear, explicit understanding of evidence-based argumentation.

4. Discussion

The findings of the present study indicate that evidence evaluation is extremely difficult for adolescents, and even adults. On all performance measures (rating of arguments, correlation of ranking with ideal ranking, and choice of best argument) only the subgroup of the 22-year-old university students showed competence. In the rating task, 65% of the university students were able to differentiate between the different levels of arguments, whereas in the 15-year-olds only a small minority of the participants passed the task. Choice of the best argument seemed to be easier, but the justifications offered for the choices again showed that only university students were clearly able to master the task (without reaching ceiling, however). Thus, the present findings indicate that argument evaluation in an informal reasoning task poses major difficulty for laypersons, with developmental change occurring late, between adolescence and adulthood, and mastery being restricted to participants with higher education. Secondly, the theoretically postulated hierarchy of levels of argument was reflected in empirical judgments in the group of university students.

The results of the study are consistent with findings of studies on argument generation skills showing that children's, adolescents' and even some adults' metaconceptual understanding of the relation between theory and evidence in informal reasoning is severely deficient. The evidence evaluation inventory presented here was developed to test participants' ability to differentiate theories from evidence under reduced task demands. While argument generation requires adequate domain-knowledge and verbal skills, the task presented in our inventory only required participants to read and remember a series of arguments. Thus, one of the obstacles to argument production, laypersons' inability to imagine themselves in the position of being able to gather relevant evidence at all, was removed. Moreover, the fact that a series of different arguments to support a theory were presented should facilitate the cognitive decoupling of theory and evidence. Rather than being in the grips of (their own) theory, participants were in the position of a third person "objectively" judging the quality of different attempts to support the theory. Despite these facilitating task conditions, however, the findings were remarkably similar to those obtained in studies of argument generation.

Only in young adults with higher educational qualification a majority consistently showed competence both in rating arguments of different quality and in choosing the best argument out of several arguments. Moreover, the justifications for choosing arguments showed that only the university students had an explicit understanding of the quality of evidence. Young adults with a vocational career generally performed significantly worse than university students and mirrored the pattern obtained in high-school students. In the rating task, the 15-year olds performed at the same level when judging Level 2 and Level 3 arguments as the young adults.

The 22-year-olds from the lower educational track did, however, outperform the high-school students on the choice and justification tasks. Thus, it might be argued that basic competencies in evaluating evidence are acquired in early adulthood across levels of education, although participants from a lower educational background need task support to demonstrate their understanding. One possible interpretation of the present findings is that skills to differentiate theory and evidence are not acquired before early adulthood, with related tasks posing a genuine conceptual problem for adolescents (and even for adults with a lower educational background).

However, it could also be argued that the argument evaluation task implied demands on working memory and text processing that may have been difficult to meet for the younger age group. The abilities of the ninth graders may have been underestimated by testing them with the same materials as the adults. Another possible explanation is that the onset of relativistic thinking in adolescence may lead ninth graders to adopt a multiplist epistemological stance (the multiplist level is characterized by the understanding that conflicting representations of the same event can be a product of interpretive mental processes that vary across individuals), which may end up in a conviction that "anything goes", "all is a matter of opinion", and "there are no good or bad arguments" (Kuhn & Weinstock, 2002).

Thus, further research is needed to investigate whether a nascent understanding of the validity of evidence can be shown in young adolescents when using supportive, flexible procedures reducing task demands. Another fruitful approach may be the use of training paradigms.

In sum, the present findings suggest that theory-evidence differentiation in informal reasoning poses a genuine conceptual problem to laypersons, indicating that the deficits in laypersons' reasoning found for argument generation tasks (Kuhn, 1991) are stable across task formats. Further research is necessary to specify the exact nature of this conceptual problem. Is it the notion of "testing" an idea or theory that is difficult to grasp per se or is it the

specific relation of empirical data to theoretical claims that people find hard to evaluate?

The present findings underscore the relation between evidence-evaluation skills and educational level. The magnitude of the difference between the lower track and the higher track groups was substantial. These differences between educational levels may develop early. In a longitudinal study on scientific reasoning, Bullock, Sodian, and Koerber (2009) found that individual differences in understanding scientific experimentation were remarkably stable from childhood through adolescence to adulthood, and could not be attributed to differential effects of schooling. Rather, the individual differences observed in fourth grade (i.e., before children were assigned to different educational tracks) persisted into adulthood. Thus, children with lower learning abilities start secondary school with a gross delay in scientific (and probably also informal) reasoning abilities. Although they show developmental progress, they do not catch up: At the age of 15 years, they perform at about the level of the 11-year-olds who were assigned to the college-bound educational track.

Reasoning abilities, such as evidence evaluation, are crucial to participating in everyday discourse in a modern democratic society. Yet, there is little effort to systematically teach such skills in secondary school curricula. The present findings demonstrate the necessity of systematic instruction, especially for lower-track students (see Kuhn, 2005). Abilities to differentiate between good and bad arguments, to identify high-quality evidence, and to evaluate the quality of evidence, are highly relevant for participation in a democratic society. People are confronted with theories (e.g., about social, economic, and political phenomena) and various types of related evidence which they are requested to evaluate in order to develop opinions and to make decisions. These opinions and decisions often have important implications for their personal life, as well as for the development of the society at large.

Thus, the educational challenge is one of reinforcing and strengthening skills already present in at least implicit forms. There is ample evidence from theory of mind research and recent research on scientific reasoning skills that even young elementary school children possess the basic ability to bring evidence to bear on a claim (Bullock & Ziegler, 1999; Sodian, Zaitchik, & Carey, 1991). There is a need for actively engaging students in thinking in order to promote insight into argumentation such as knowledge about the structure of arguments and about the characteristics of a good argument, as well as the enhancement of argumentation skill via writing (Voss & Means, 1991), teaching metacognitive knowledge about argumentation (Perkins, Farady, & Bushey, 1991), and using authentic problems from everyday life with relevance to the learners' situation.

Learning is not simply the accumulation of facts about how the world is. Learning involves the construction of theories that provide explanations for how the world may be. Science often progresses through dispute, conflict, and argumentation rather than through general agreement (e.g., Kuhn, 1962; Latour & Woolgar, 1986). Thus, arguments concerning the appropriateness of experimental design, the interpretation of evidence, and the validity of knowledge claims are at the heart of science, and are central to the everyday discourse of scientists. Scientists engage in argumentation and it is through this process of argumentation within the scientific community that quality control in science is maintained (T.S. Kuhn, 1962). Lessons involving argument will require children to externalize their thinking. Such externalization requires a move from the *intra*-psychological plane, and rhetorical argument, to the *inter*-psychological and dialogic argument (Vygotsky, 1978). When children engage in successful argumentation, and support each other in high-quality argument, the interaction between the personal and the social dimensions promotes reflexivity, appropriation, and the development of knowledge, beliefs, and values. To grasp the connection between evidence and theory is to sharpen children's ability to think critically in a scientific context, preventing them from becoming blinded by unwarranted commitments. From sociocultural perspectives on cognition, argumentation is a critical tool for science learning since it enables within learners the appropriation of community practices including scientific discourse.

Appendix

1. Problems and candidate causes presented in the interview

1.1 Aggression
Problem: What is the cause of increasing aggression and violence among children and adolescents?
Candidate cause**:** Children and adolescents become aggressive and violent because of the continuous consumption of media with violent content, like video games and horror films.

1.2 Birthrate
Problem: What is the cause of the decline of the birthrate in Germany?
Candidate cause: Women decide against having children because the role of a housewife and mother is less attractive nowadays than it used to be in former times.

1.3 Math

Problem: Why do some students perform well in mathematics and some do not?
Candidate cause: Math is a matter of giftedness and therefore genetically determined. Good performance in math is in your genes.

1.4 Smoking
Problem: What is the cause of the increasing consumption of cigarettes among children and adolescents?
Candidate cause: If babies or toddlers are exposed to the smoke of cigarettes of their parents, they become addicted to nicotine through passive smoking. This early addiction to nicotine makes them prone to start smoking once they are teenagers.

1.5 Overweight
Problem: What is the cause of being overweight?
Candidate cause: Individuals can not change their unhealthy diet to reduce weight on a sustained basis.

1.6 Success in Occupation
Problem: Why do some people succeed in occupation and career and some do not?
Candidate cause: Success in occupation depends on the social background of the family. Sons and daughters of university graduates become university graduates themselves. Parents from society circles have more influence and can pave the way for their children more effectively than parents from the working class can.

2. Level of validity of the presented evidence in the interview

Level 1: *No differentiation of theory and evidence*
The argument consists of a further description of the theory by elaborating mechanisms consistent with the theory or by specifying new instances of the presumed causal chain. No genuine evidence is provided for the theory, nor is there any reflection on possible ways of generating relevant evidence.

Arguments at level 1 also included examples of two standard reasoning fallacies: The "ad populum" argument appeals to a systematic feeling of group solidarity in order to prove the point ("If most of the people believe this, then so do I"). In the "ad ignorantiam" argument a conclusion about the truth of a proposition is derived from the fact that the proposition is not known to be false.

Level 2: *Partial differentiation of theory and evidence*

Observations and findings are brought to bear on a causal theory; however, it is unclear why and how these pieces of evidence are supposed to support the theory. For example, single cases of co-occurrence of cause and effect are cited, or observations that are not clearly related to the theory.

Level 3: Simple differentiation of theory and evidence
There is a clear understanding of the notion of an empirical test for a theory. However, evidence supporting simple beliefs, rather than more complex belief systems or theories is generated.

Level 4: Full differentiation of theory and evidence
Arguments are based on an understanding of the target theory and its relation to alternative theoretical explanations. There is a clear understanding that findings of an empirical test can be interpreted in different ways depending on theoretical viewpoints. Thus, a series of investigations is outlined to rule out alternative interpretations. For example, a set of empirical tests is laid out to test specific hypotheses derived from the theory and to rule out alternative interpretations of a set of findings.

References

Barchfeld, P., & Sodian, B. (in prep.). *The development of argumentation skills from childhood to young adulthood.*

Bullock, M., Sodian, B., & Koerber, S. (2009). Doing experiments and understanding science: Development of scientific reasoning from childhood to adulthood. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood: Findings from a 20-year longitudinal study* (pp. 172-198). New York: Psychology Press.

Bullock, M. & Ziegler, A. (1999). Scientific reasoning: Developmental and individual differences. In F. E. Weinert & W. Schneider (Eds.). *Individual Development from 3 to 12. Findings from the Munich Longitudinal Study* (pp. 55–61). Cambridge: University Press.

Brem, S.K., & Rips, L.J. (2000). Explanation and evidence in informal argument. *Cognitive Science, 24*, 573-604.

Carey, S., Evans, R., Honda, M., Jay, E., Unger, M. (1989*).* "An experiment is when you try it and see if it works": A study of grade 7 students´ understanding of the construction of scientific knowledge. *International Journal of Science Education, 11,* 514-529.

Glassner, A., Weinstock, M., & Neuman, Y. (2005). Pupils' evaluation and generation of evidence and explanation in argumen-

tation. *British Journal of Educational Psychology, 75,* 105-118.

Kuhn, D. (1991). *The skills of argument.* New York: Cambridge University Press.

Kuhn, D. (1992). Thinking as argument. *Harvard-Educational-Review, 62*, 155-178.

Kuhn, D. (2005). *Education for thinking*. Cambridge, MA: Harvard University Press.

Kuhn, D., & Felton, M. (2000, January). *Developing appreciation of the relevance of evidence to argument.* Paper presented at the Winter Conferences on Discourse, Text and Cognition, Jackson Hole, W.Y.

Kuhn, D., & Franklin, S. (2006). The second decade: What develops (and how). In D. Kuhn & R.S. Siegler (Eds.). *Handbook of child psychology*. (vol.2.). Cognition, perception, and language (pp. 953-993). New York: Wiley.

Kuhn, D., & Pearsall, S. (2000). Developmental origins of scientific thinking. *Journal of Cognition and Development, 1*, 113-129.

Kuhn, D. & Weinstock, M. (2002). What is epistemological thinking and why does it matter. In B.K. Hofer & P.R. Pintrich (Eds.). Personal Epistemology (pp121-145). Mahwah: Laurence Erlbaum.

Kuhn, D., Weinstock, M., & Flaton, R. (1994). How well do jurors reason? Competence dimensions of individual variation in a juror reasoning task. *Psychological Science, 5,* 289-296.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University Press.

Latour, B., & Woolgar, S. (1986). *Laboratory Life: The Construction of Scientific Facts*., Princeton: Princeton University Press.

Leadbeater, B., & Kuhn, D. (1989). Interpreting discrepant narratives: Hermeneutics and adult cognition. In J. Sinnott (Ed.), *Everyday problem solving: Theory and Application* (pp. 175-190). New York: Praeger.

Means, M.L., & Voss, J.F. (1996). Who reasons well? Two studies of informal reasoning of different grade, ability, and knowledge levels. *Cognition and Instruction, 14*, 139-178.

Perkins, D.N., Farady, M., Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J.W. Segal, J.F. Voss, D.N. Perkins (Eds.). *Informal reasoning and Education* (pp. 83-105). Hillsdale: Erlbaum.

Sá, W., Kelley, C.N., Ho, C., Stanovich, K.E. (2004). Thinking about personal theories: Individual differences in the coordination of theory and evidence. *Personality and Individual Differences, 38,* 1149 – 1161.

Sodian, B., Zaitchik, D. & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62,* 753-766.

Voss, J.F. and Means, M.L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction, 1,* 337-350.

Vygotsky, L.S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Weinert, F.E., & Schneider, W. (1999*). Individual development from 3 to 12. Findings from the Munich longitudinal study*. New York: Cambridge University Press.

Weinstock, M., & Cronin, A. (2003). The everyday production of knowledge: Individual differences in epistemological understanding and juror-reasoning skill. *Applied Cognitive Psychology ,17,* 161-181.