Thinking & Reasoning

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ptar20

The stability of syllogistic reasoning performance over time

Hannah Dames, Karl Christoph Klauer & Marco Ragni

To cite this article: Hannah Dames, Karl Christoph Klauer & Marco Ragni (2021): The stability of syllogistic reasoning performance over time, Thinking & Reasoning, DOI: 10.1080/13546783.2021.1992012

To link to this article: https://doi.org/10.1080/13546783.2021.1992012

Published online: 28 Oct 2021.

THINKING & REASONING
https://doi.org/10.1080/13546783.2021.1992012

Routledge
Taylor & Francis Group

Check for updates

The stability of syllogistic reasoning performance over time

Hannah Dames[a,b], Karl Christoph Klauer[c] and Marco Ragni[a,d]

[a]Department of Computer Science, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany; [b]Department of Psychology, University of Zurich, Zurich, Switzerland; [c]Department of Psychology, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany; [d]Danish Institute of Advanced Studies, South Denmark University, Odense, Denmark

ABSTRACT
How individuals reason deductively has concerned researchers for many years. Yet, it is still unclear whether, and if so how, participants' reasoning performance changes over time. In two test sessions one week apart, we examined how the syllogistic reasoning performance of 100 participants changed within and between sessions. Participants' reasoning performance increased during the first session. A week later, they started off at the same level of reasoning performance but did not further improve. The reported performance gains were only found for logically valid, but not for invalid syllogisms indicating a bias against responding that 'no valid conclusion' follows from the premises. Importantly, we demonstrate that participants substantially varied in the strength of the temporal performance changes and explored how individual characteristics, such as participants' personality and cognitive ability, relate to these interindividual differences. Together, our findings contradict common assumptions that reasoning performance only reflects a stable inherent ability.

ARTICLE HISTORY Received February 27, 2021; Accepted October 6, 2021

KEYWORDS Reasoning; retest effect; interindividual differences; stability

Introduction

Individuals differ in their ability to reason logically (e.g., Frey et al., 2018; Khemlani & Johnson-Laird, 2016; Newstead et al., 2004; Stanovich & West, 2000). The question of why some individuals are able to draw a logically correct conclusion from given information while others are not, has received much attention (e.g., Galotti et al., 1986; Newstead et al., 2004; Stanovich & West, 2000; Svedholm-Häkkinen, 2015). Most of the studies

CONTACT Hannah Dames damesh@cs.uni-freiburg.de Department of Psychology, University of Zurich, Binzmühlestrasse 14/22, Zurich, 8050, Switzerland.

investigating individual differences in reasoning performance assume that the employed tasks measure an individual's ability to reason logically. Yet, it is still unknown whether participants' performance in those tasks measured within an experimental session reflects a stable reasoning construct or corresponding stable cognitive mechanisms. That is, we do not yet know whether our measurements reflect stable reasoning processes, and/or a stable latent ability, or whether they are also affected by other variables such as changes in strategies over time and the reduction of construct-irrelevant factors.

The assumption that an individuals' reasoning behaviour is based on a stable, latent capacity (see Rips, 1994; Schaeken et al., 2000) can be questioned: Studies on the usage of reasoning strategies (e.g., Bucciarelli & Johnson-Laird, 1999; Roberts & Newton, 2003) as well as the literature on retest effects for cognitive ability tests in general (see Lievens et al., 2007) — suggest that individuals' reasoning performance may change over time. Of note, individuals may in particular differ in the extent of such temporal changes in performance. In a series of three experimental sessions, we thus investigated a) whether individuals' deductive reasoning performance changes over time within as well as between experimental sessions, b) whether individuals substantially differ from each other in these temporal changes, and c) to what extent individuals' characteristics explain these interindividual differences. For this purpose, we employed a traditional syllogistic reasoning task as well as a test battery consisting of a variety of cognitive ability tests as well as personality assessments.

The syllogistic reasoning task

Syllogistic reasoning is one of the core domains in human reasoning research (for a review see Khemlani & Johnson-Laird, 2012). A syllogism consists of two premises, where each premise contains one of four quantifiers *All* (abbreviated by A), *Some* (I), *Some..not* (O), and *None* (E), and two terms (denoted in the following by A, B, and C). Consider the following example of a syllogism:

All A are B.

Some B are C.

What, if anything, follows?

Hence, the task is to derive a logically correct conclusion (i.e., a "valid" conclusion) or to state that logically "nothing follows" (i.e., "no valid conclusion" NVC for short). For the example above, participants often infer that "Some A are C" (Khemlani & Johnson-Laird, 2012). However, the logical valid response is that nothing follows. By rearranging the order of terms four different figures can be formed (Khemlani & Johnson-Laird, 2012). In total, given two premises and the four quantifiers, there are 64 distinct structural forms of syllogisms. In the syllogistic reasoning task, participants are often instructed to generate a quantified answer using one of the four quantifiers (A, I, O, E) about the two sets A and C connected by the middle term B or to conclude that no logically valid conclusion follows. In different versions of the task, participants are instructed to generate a conclusion, to choose a response from a set of given conclusions, or to evaluate a given conclusion. Here, we employed a constrained generation task as elaborated on below.

The capacity to reason and the involved cognitive processes have been the focus of psychological research for many years (e.g., Johnson-Laird & Byrne, 1991; Störring, 1908) and many theories on cognitive processes underlying reasoning have emerged since (see Khemlani & Johnson-Laird, 2012). Whereas some theories of reasoning postulate that individuals use formal rules of inference akin to those of logic (e.g., Rips, 1994), other researchers suggest that people simply use heuristics based on the surface characteristics of a problem (e.g., Chater & Oaksford, 1999; Woodworth & Sells, 1935) or build mental models based on the given information and carry out operations on these representations (Johnson-Laird, 1980). Although most cognitive theories seem to predict reasoning performance to some degree, their ability to predict individual human responses is somewhat limited (see Brand et al., 2019; Riesterer et al., 2018; Riesterer, Brand, Dames, et al., 2020). Yet, the lack of predictive performance of these models cannot entirely be attributed to random noise in the data (Riesterer, Brand, & Ragni, 2020). Such results may indicate two important issues: first, in addition to the assumed cognitive reasoning mechanisms proposed by the theories, other variables may influence individuals while solving a reasoning task. Second, people may differ highly in their cognitive reasoning strategies/processes themselves or in the quality of such strategies/processes (e.g., the extent to which participants engage in a certain process). The question of how people differ when reasoning is still controversial and highly discussed (e.g., Frey et al., 2018; Galotti et al., 1986; Khemlani & Johnson-Laird, 2016; Newstead et al., 2004; Stanovich & West, 2000). Variables that have been found to be associated with individual differences in reasoning range from differences in intelligence, working memory, or other cognitive abilities (Süß et al., 2002), differences in trait characteristics – such as generally preferring logic over intuition or vice versa (e.g., Svedholm-Häkkinen, 2015) to metacognitive processes (Ackerman & Thompson, 2018). In the current study we consider how reasoning performance changes over time and whether individual characteristics relate to differences in such temporal changes in reasoning performance.

If dynamic adaptation by an individual reasoner to a reasoning task over time exists, cognitive models would benefit from explicitly integrating them. Considering such temporal changes over time is important, because – as Evans (2011) points out – in contrast to many existing theories, not only cognitive capacity influences task performance but also other, more dynamic factors, metacognitive processes (e.g., metacognitive monitoring processes that accompany the reasoning process), and strategies in the form of acquired rules and procedures. Relatedly, various studies in psychometric research demonstrated that construct-irrelevant and temporally variable factors distort individuals' performance on, for instance, intelligence tests (Lievens et al., 2007; Matton et al., 2009). However, those effects are rarely taken into consideration when investigating syllogistic reasoning tasks. We assume that – in addition to the aforementioned factors – such variables influence reasoning behavior and result in performance changes throughout the time-course of experimental sessions. Given the great length of many syllogistic reasoning experiments (e.g., testing all 64 syllogisms in one session which can take more than one hour), neglecting such influences can become highly problematic, as strategies and motivational factors may unfold throughout the experiment in different ways for different individuals. Consequently, considering reasoning trajectories over time could help to explain some of the substantial variance we find in syllogistic reasoning data.

The present research: the flexibility and stability of reasoning performance over time

In this study, we investigate the extent of, and individual differences in, changes in syllogistic reasoning performance over time. In order to achieve this goal, we conducted a study consisting of three experimental sessions. In the first one, individual characteristics such as an individual's cognitive ability and personality were measured. As in this first session only individual characteristics were measured, but no syllogisms administered, we refer to this pre-assessment as Session 0. In the second session (referred to as Session 1), participants performed the syllogistic reasoning task consisting of all 64 syllogisms and completed the Cognitive Reflection Test. In a third session (Session 2), participants again responded to all 64 syllogisms and completed the Raven test. All sessions were separated by one week. This design allowed us to investigate how individuals differ in their reasoning ability over time both within an experimental session as well as between two test sessions as a function of participants' cognitive ability and personality. Two main classes of factors have been argued to result in a change, specifically an improvement, of reasoning performance over time:

(1) the development and usage of test-specific strategies and (2) the reduction of construct-irrelevant influences.

Retest effects in psychometric research: the influence of Construct-Irrelevant factors

The question of how individuals improve in cognitive ability test scores as a result of retaking the same cognitive ability test under comparable conditions has been the subject of classical research on retest effects (Lievens et al., 2007), often also referred to as *testing* (Roediger & Butler, 2011) or *practice effects* (Hausknecht et al., 2007; for an overview see Scharfen et al., 2018). The term retest effect thus refers to the frequently observed increase in test scores across repeated administration of the same or a similar test (Lievens et al., 2007). Retest studies using all 64 syllogistic problems are, however, virtually non-existent, with the exception of one study (based on a sample from 1978, reported in Ragni et al., 2018) with 20 participants. Studies in the field of psychometric research propose three main factors underlying the retest effect in cognitive ability tests (for an overview, see Lievens et al., 2007): the enhancement of the latent construct itself measured by the test – a view that is strongly contested for cognitive ability tests (e.g., Hausknecht et al., 2002; Lievens et al., 2007), a reduction in construct-irrelevant, distorting factors, and last, the development of test-specific strategies as just discussed.

Similar to cognitive ability tests, syllogistic reasoning experiments are not conducted in a situational vacuum and such experiments are essentially a testing situation. As a consequence, we can assume that factors such as the unfamiliarity with the testing situation and test anxiety can impact participants' reasoning behavior (see also Reeve et al., 2009 for an example how construct-irrelevant factors such as test anxiety and test familiarity can influence the criterion-related validity of cognitive ability tests). Evidence suggests that these factors diminish participants' test performance in various cognitive ability tests (Eysenck et al., 2007; Ng & Lee, 2015), and that their impact declines when the participants are retested (for an overview, see Scharfen et al., 2018). We thus predict that over time the influence of construct-irrelevant, distorting factors also decreases, leading to improvements in performance.

The use of strategies in syllogistic reasoning tasks

Deductive reasoning has been claimed to be a stable capacity (see Rips, 1994; Schaeken et al., 2000). Evidence suggests, however, that individuals sometimes learn to reason by discovering new strategies while performing the task (for an overview of individual differences in reasoning strategy selection and availability, see Roberts & Newton, 2003). Reasoners do not only use a variety of strategies (between and within individuals), individuals also seem to change these strategies throughout an experimental session (Bucciarelli & Johnson-Laird, 1999; Roberts & Newton, 2003).

There seems to be no clear definition on what counts as a strategy in reasoning tasks. On a rather broad and abstract level, strategies can refer to the use and manipulation of visuospatial or verbal-propositional representations (Bacon et al., 2003): Whereas spatial reasoners produce an explicit representation of relationships between terms and premises, verbal reasoners manipulate information in its abstract form by, for instance, switching the terms in the premises. According to Bacon et al. (2003), the aforementioned verbal strategies however also include the use of simple rules which define conclusions as associated with particular combinations of quantifiers, for instance, whenever given *All* and *None*, "*No valid conclusion*" follows. Task-specific shortcut strategies of this kind can readily account for changes in reasoning performance over time. To give an example, the two-some rule for syllogistic reasoning (Galotti et al., 1986) describes how some participants spontaneously seem to develop a rule, where, for any given syllogism, when the word *some* appears twice, there is never a valid conclusion. The application of the two-some rule is a good example for how the identification of a strategy can lead to massive gains in speed and accuracy for the syllogisms to which it can be applied. Note that the two-some rule, although a verbal strategy, can be detected by reasoners using visuospatial strategies on problems with two occurrences of "some".

Interestingly, some strategies seem to be acquired over the course of one experimental session as questionnaire responses suggested that applied strategies did not result from earlier experiences, such as skills learned in school (Bacon et al., 2003). It is therefore reasonable to assume that some participants can develop, identify, and select efficient strategies in the course of an experiment. These participants should then improve over time (see the two-some rule).

Taken together, participants' responses to reasoning problems may improve over time due to the reduction of distorting factors and the increased use of helpful strategies. We propose that this applies to participants' reasoning performance (1) within a session and (2) between two test administrations and that (3) individuals substantially differ from one another in these performance changes. We speculated that the development of test-specific strategies would be associated with cognitive-ability and thinking-disposition measures whereas the reduction of distorting factors such as test anxiety might be associated with personality traits, motivating the use of measures of cognitive abilities as well as personality traits in this study.

Increase in reasoning performance within an experimental session

The selection of effective test-specific strategies should result in an improvement within an experimental session (and thus over time) even without feedback. In addition, effects of practice and repetition may occur within one experimental session (e.g., based on habituation) due to a reduction of distorting and construct-irrelevant factors (Freund & Holling, 2011; Lievens et al., 2007; Matton et al., 2009). This, in turn, could lead to an increase in reasoning performance. However, at the same time, we are well aware that effects of fatigue or a loss of motivation may counteract typical practice effects. Yet, assuming participants are – on average – motivated to fully engage in the task throughout one session, we assume that reasoning performance increases as the trial sequence proceeds for most of our participants.

Increase in reasoning performance between two test sessions

We expect that the improvement of reasoning performance within an experimental session should be stable over time (i.e., between two test sessions). The same key variables causing an improvement in the first session apply for the second test administration: Construct-irrelevant factors that were reduced already during the first test session, should still be diminished in the retest. Strategies successfully developed and applied in the first test session should again facilitate a better reasoning performance when being retested. Empirical evidence for the development of such strategies can for instance be found in a study on the Raven Matrices test conducted by Hayes et al. (2015). Here, the authors demonstrated that procedural knowledge tacitly acquired during training can later be utilized at post-test. These considerations result in clear predictions regarding a potential retest effect: First, we expected that participants' reasoning performance increased between two test sessions (retest effect).

However, with respect to the use of strategies, we argued that participants' performance should already improve during the first test administration. We assume that this improvement is still existent (e.g., if strategies are still available) in Session 2, meaning that participants should demonstrate a similar likelihood to respond correctly at Trial 1 in Session 2 as at Trial 64 in Session 1. Consequently, we predicted that a potential retest effect would be greater when compared for low trial numbers between sessions and that it would be reduced for later trials. This is in line with the assumption that although performance may still improve during the second test session, the slope for the influence of increasing trial number should be reduced in the second session.

Finally, we propose that individuals do not only differ in their average reasoning performance, but also in the magnitude of performance gains over time. That is, we predict that the changes in reasoning performance within and between sessions differ between participants. To empirically test this prediction, we employ a hierarchical modeling approach that accounts for those interindividual differences.

Second, to further substantiate the existence of interindividual differences in temporal performance changes, we administer a test battery consisting of a variety of cognitive ability tests as well as personality assessments[1]. This allows us to investigate whether individual differences in an effect of time (trial number and session) are associated with participants' cognitive ability and/ or personality – a question we aim to investigate exploratively, that is without prior hypotheses.

Methods

Participants

The initial sample consisted of $N = 114$ participants. Participants' mother tongue was German. We excluded participants who did not complete the full study (drop-outs: $n = 7$). Two participants had to be excluded from the study due to misunderstanding the task and resulting non-sensical answers (both responded "no valid conclusion" for each problem). Furthermore, to ensure that participants were not trained in logics, we excluded participants with an educational background in mathematics, philosophy with logic-courses, and computer science ($n = 5$). The final sample consisted of $n = 100$ participants (69% female, 30% male, 1% other; $M_{age} = 25.23$, $SD_{age} = 5.77$). Most of the participants (92%) had graduated from high-school (German: "Abitur") and 63% of the participants were enrolled in a university program or in another form of training. The study was carried out in strict accordance with the ethical principles as formulated in the WMA Declaration of Helsinki and the research standards set by the German Psychological Society.[2]

[1] For the assessment of cognitive abilities, we chose constructs that correlated with an individual's ability to reason in previous work: working memory capacity (e.g., Copeland & Radvansky, 2004; Süß et al., 2002), intelligence (e.g., Süß et al., 2002), and a person's disposition for reflective thinking (e.g., Svedholm-Häkkinen, 2015; Toplak et al., 2014). There are only a few studies that have examined the role of personality in reasoning (Brase et al., 2019). At the same time, there has been an increase in work on the link between personality and cognitive abilities in general (especially intelligence, e.g., Carretta & Ree, 2018). On this basis, we selected the measures for which we assumed that similar relations would unfold for the relation between personality and reasoning performance: Previous studies reported negative relations between some factors of the Big Five and cognitive abilities or intelligence, notably for extraversion, neuroticism, and conscientiousness (e.g., Carretta & Ree, 2018; Moutafi et al., 2003; Moutafi et al., 2004; Moutafi et al., 2006; Rammstedt et al., 2016). Furthermore, individuals with a high Need for Cognition tend to engage in tasks that are time-consuming and require effortful thinking. Hence, we also administered participants' Need for Cognition (Frederick, 2005).

[2] If research objectives do not involve issues regulated by law (e.g., the German Medicine Act [Arzneimittelgesetz, AMG], the Medical Devices Act [Medizinproduktegesetz, MGP], the Stem Cell Research Act [Stammzellenforschungsgesetz, StFG] or the Medical Association's Professional Code of Conduct [Berufsordnung der Ärzte]), then no ethics approval is required for social science research in Germany. Our study had no such objectives, and therefore, no IRB approval or waiver of permission was sought for it.

Measures

All materials were presented in German. Here, we briefly describe the instruments used to measure participants' individual characteristics. More detailed descriptions are available online[3].

The Big Five personality traits were assessed with the German Big-Five-Inventory-SOEP (BFI-S, Gerlitz & Schupp, 2005) comprising 15 items, with three items for each Big Five dimension extraversion, agreeableness, conscientiousness, neuroticism, and openness (agreeableness was not considered here). Considering the low number of items in the current study, Cronbach's alpha values were acceptable (Neuroticism: $\alpha = .82$, Extraversion: $\alpha = .83$, Openness: $\alpha = .76$, Conscientiousness: $\alpha = .65$) and were comparable or even above the reported range of Cronbach's alpha scores for that instrument (see Gerlitz & Schupp, 2005).

To measure the individual's tendency to engage in and enjoy reflective, complex, and challenging thinking, the German short Version (Bless et al., 1994) of the Need for Cognition scale was administered (16 items, Cacioppo & Petty, 1982). In the current study, internal consistency for the Need for Cognition scale was good ($\alpha = .89$).

Visuo-spatial working memory was assessed using a computerized version of the Corsi block-tapping task (Corsi, 1973; Milner, 1971). The mean of participant's estimated highest forward and backward span (max 9) was used as a measure for visuo-spatial working memory capacity.

Verbal working memory capacity was assessed by a complex span task using a modified version of the Turner and Engle (1989) operation span test following the procedure and item sets of an open-source version (Stone & Towse, 2015). We used the proportion correct method of scoring (referred to as prop-score, which is the proportion of items that a participant recalled in the correct serial position during the task).

- For the assessment of participants' general cognitive ability, we used a non-verbal task, the short form of the Advanced Progressive Matrices (RAPM, Bors & Stokes, 1998, referred to as the Raven test in this study. Here, Cronbach's alpha was acceptable: $\alpha = .75$).

To measure interindividual differences in the individuals' intuitive–analytic cognitive styles, the extended seven-item Cognitive Reflection Test consisting of the original three items (Frederick, 2005) and four additional items (Toplak et al., 2014) were used. In our study, Cronbach's alpha for the Cognitive Reflection Test was acceptable: $\alpha = .74$.

[3] https://osf.io/x3wvf/

Note, that this study also administered a short conditional reasoning test within the first session and Self-efficacy expectations (German General Self-Efficacy Short Scale, German: ASKU; Beierlein et al., 2012; Beierlein et al., 2013), which we do not consider in this study.

The syllogistic reasoning task

Material

In the syllogistic reasoning tasks, participants had to generate a conclusion for all 64 distinct categorical syllogism problems. Each problem consisted of two premises and asked for a conclusion that can be drawn based on these premises (see the introduction for an example) using a production task design: Participants were asked to generate a conclusion corresponding to the typical syllogistic response format (see Procedure). For this work, a novel German set of syllogisms items was created (see Appendix A, Table A1). The content of the syllogisms (placeholders A, B, and C) was randomly drawn from three lists: names of professions (e.g., Jurist [lawyer]), sports (e.g., Boxer [boxer]), and hobbies/ musicians (e.g., Gitarrist [guitarist]). The chosen German occupations, sports, and hobbies/musicians were frequent and well-known in order to ensure familiarity with the groups of interests. Different content was used in each syllogism, and the terms A, B, and C of a given syllogism came from different lists to ensure the believability of the statements (e.g., people are not likely to work on two types of job such as being a lawyer and a secretary at the same time). Under these constraints, the combination of terms within a syllogism was randomly selected for each item and participant.

Procedure

Participants received standard syllogistic reasoning instructions stressing that the premises should be assumed to be true, and that a conclusion should be drawn only if it followed logically given the two premises. Before the start of the task, participants were familiarized with the presentation format using an example item. Participants were instructed to draw a conclusion linking the two end terms of the syllogism following the form "all … are … ", "no … are … ", "some … are … ", "some .. are not … ", or to indicate that no valid conclusion followed. Importantly, they were not given any information on how to solve syllogisms. Syllogisms were presented one at a time in different random orders for each participant and each session.

The premises of a syllogism were simultaneously presented on the screen center. Participants were instructed to press the spacebar as soon as they decided on an answer that can be derived from the two premises.

Upon pressing the spacebar, a box appeared in which participants were told to type their response. On screen, participants were reminded that their answer (other than the answer that no valid conclusion follows) should contain one of the quantifiers-"some," "all," "no," or "some .. not"- and that it should link the two end terms of the premises. This was accomplished by presenting all possible response forms at the bottom of the screen under the response box (see Appendix B, Figure A1). The order of the response-format options was random for each trial in order to reduce confounding influences of response pattern effects. Premises remained on screen while participants entered the conclusion using the keyboard. To ensure data-quality response restrictions were applied. That is, when the participant's typed response did not follow the response format (e.g., contained the middle term), participants received a message reminding them about the response format and were asked to change the answer accordingly. After responding to a syllogism, participants proceeded to the next syllogism. Participants did not receive any feedback on their responses' correctness. Conclusions had to be determined using no additional materials (i.e., no paper or pencil was allowed).

Experimental set-up

All three sessions took place in a soundproof laboratory room (one participant per room) in front of a computer screen. Responses were collected via the keyboard and a mouse.

Overall procedure

The experiment consisted of three sessions, each one week apart from the next (see Figure 1). All three sessions were computer based. The participation timeslots were allotted based on strict requirements. Participants could only book timeslots that were (1) one week apart from one another, (2) on the exact same weekday, and (3) around the same time of the day (i.e., ±1 hour). This was done to ensure that possible time-of-day and day-of-week effects were controlled within a participant. For all experimental sessions, informed consent was given. Participants were told that data were recorded in anonymous format and that they could quit the experiment at any time, but if they did, they would not be able to participate in subsequent sessions. They were then led to the laboratory room and all following instructions were computer-based. Participants produced a personal code that was saved in order to match different datasets of each participants. Throughout all sessions, participants were able to read instructions for as long as they wanted. Figure 1 provides an overview of the measures

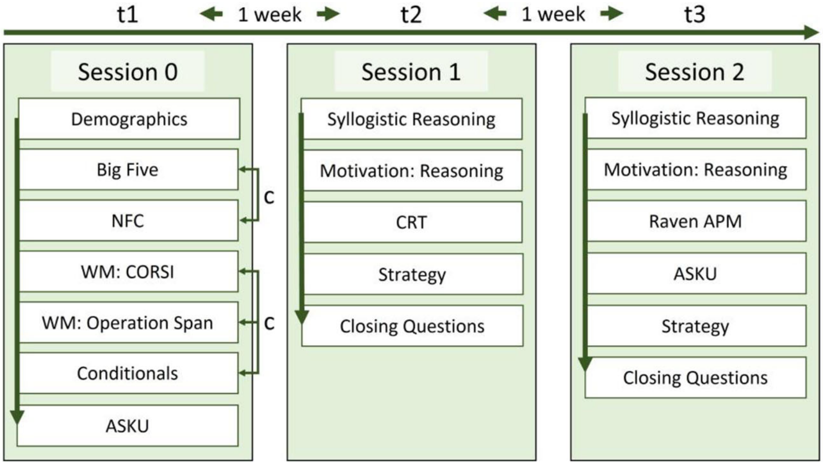


Figure 1. Overview of the measures obtained per session.

Note. NFC = Need for Cognition, WM = working memory, ASKU = Allgemeine Selbstwirksamkeit Kurzskala (referring to the German General Self-Efficacy Short Scale). APM = Advanced Progressive Matrices, CRT = Cognitive Reflection Test. c = counterbalanced order.

collected during each session. Participants were told that their payment would depend on their performance of the tasks. At the end of the experiment, participants were paid on the basis of the time they had invested in the experiment with a rate of 7.50 EUR per hour.

Session 0. First, we selected participants' demographic data (age, educational background, and language skills). Then, we administered the BFI-S, the Need for Cognition scale, the conditional reasoning task, the CORSI Block-Tapping Test, the operation span test, and the German General Self-Efficacy Short Scale.

Session 1. Participants started with the syllogistic reasoning tasks. After having completed 16 and 48 out of all 64 syllogistic reasoning tasks, participants had the opportunity to take a self-paced break. After 32 out of the 64 reasoning tasks (thus, 32 problems), participants were instructed to engage in a five-minute break and to leave the laboratory rooms to take a short walk. This pause was "mandatory" for all participants in order to reduce effects of fatigue or other distorting factors. After completing all 64 syllogisms, participants answered questions regarding their guessing rate, motivation and attention while completing the syllogistic reasoning tasks. Next, they were again asked to take a mandatory break. Participants then completed the Cognitive Reflection Test. At the end of the session, participants answered general questions about their overall motivation and tiredness during the experiment. Here, they were asked whether they had used any strategy during the reasoning task and if so, they were asked to describe the selected strategies.

Session 2. Session 2 was identical to Session 1 with the exception that the Raven test and the German General Self-Efficacy Short Scale were administered at the end of Session 2.

The mean overall completion times of Session 0, 1, and 2 were, in order, around 42 minutes ($SD = 11$ minutes), 1 hour and 21 minutes ($SD = 24$ minutes), and 1 hour and 9 minutes ($SD = 18$ minutes).

Data analysis

All data and analysis scripts are available on the Open Science Framework (https://osf.io/x3wvf/). The extent of any improvement in performance over the two sessions was determined by submitting the correctness of participants' trialwise responses as a dependent variable to a generalized linear mixed model (GLMM, for an overview on mixed models see Baayen et al., 2008; Judd et al., 2012) with *validity* of a syllogism (valid $= -1$, invalid $= 1$), *trial number* (numeric, 1-64, centered and scaled), *session* (Session $1 = 0$, Session $2 = 1$), and their interactions as fixed factors, and participant as well as syllogisms as random factors. We ran GLMMs with a logistic link function for participants' responses' correctness (correct or incorrect response) using mainly the R packages lme4 (Bates et al., 2015) and afex (Singmann et al., 2017) with the maximum likelihood method and the "BOBYQA" optimizer. We obtained model predictions and graphics with the support of the ggeffects package (Lüdecke, 2018).

The data analysis procedure was as follows. First, we estimated a GLMM that consisted of the most complex random effect structure justified by the design. In case this model did not converge we reduced the random effect structure. Second, we determined the most appropriate random effects structure and thereby the final model used for the subsequent analyses (for a similar model selection procedure, see Bender et al., 2016): The procedure for selecting the model with appropriate random-effects structure is described in Appendix C. Third, using the final model that resulted from the previous operations, we then investigated the fixed effects of the model to test our hypotheses.

The *p*-values were estimated via Likelihood-Ratio-Tests and we additionally calculated the odds ratio statistic for interpretation of effect sizes (OR, see Szumilas, 2010). All continuous predictor variables were centered on their mean. As we were interested in the increase in reasoning performance between the two sessions and Session 1 served as a meaningful baseline, the contrast coefficient for *session* was intentionally set to 0 for the first session with syllogisms (Session 1), and 1 for the second session (Session 2). For all other fixed factors, sum-to-zero contrasts were used. For each model, we calculated marginal and conditional R^2 statistics, based on Nakagawa et al. (2017). The marginal R^2 considers only the variance accounted by the fixed effects, while the conditional R^2 takes both the fixed and random effects into account.

To determine whether participants reliably differed in the effect magnitude of *trial number*, *session*, and their interaction, using Likelihood-Ratio Tests, we compared different versions of the final model (obtained in the previous step) that included a by-participant random slope for the effect of interest to a model version that did not include that random slope. If a model with the by-participant random slope for the factor of interest fitted the data significantly better than a model without that random slope, this is evidence in favor of our prediction that there is reliable between-participants variability in the slope associated with the effect of *trial number*, *session*, and/or their interaction.

In a last step, we assessed how the cognitive ability measures and personality traits relate to the reported changes in reasoning performance over time by fitting five additional models for each measure: a working memory model (operation span + Corsi), a Raven model, a Cognitive Reflection model, a big five model (extraversion, neuroticism, openness, and conscientiousness), and an Need for Cognition model. For each model, we added the corresponding measure(s) and all possible interactions between a single measure and all previous fixed effects (session, trial number, validity, and their interaction) as fixed effects. All measures were z-standardized. To account for performing multiple tests when determining which individual characteristics measures were substantially associated with the effect magnitudes of *session*, *trial number*, and their interaction, we adjusted the overall alpha level of .05 using the Holm-Bonferroni Method. We separately corrected the alpha level for interactional effects between the four cognitive ability measures and the effect of *trial number* (four tests), between the cognitive ability measures and the effect of *session* (four tests), as well as between the cognitive ability measures and the *trial number x session* interaction (four tests). Likewise, alpha levels were separately adjusted for interactional effects between the four cognitive ability measures and the *validity x session* interaction, between the cognitive ability measures and the *validity x trial number* interaction, as well as between the cognitive ability measures and the *validity x trial number x session* three-way interaction. The same correction method was applied for the five personality measures (five tests each). Thus, for cognitive ability measures an observed *p*-value was compared to its corresponding adjusted alpha level for statistical inference of .0125, .0167, .025, .050 (and for personality measures: .010, .0125, .0167, .025, .05). The significance of an effect was tested in order from the smallest to largest *p*-values.

Table 1. Percentages of logically correct responses averaged for Session 1 and 2 and for valid and invalid syllogisms.

	Mean Accuracy (%)						
	Session 1				Session 2		
	overall	*valid*	*invalid*		*overall*	*valid*	*invalid*
Mean	50.1	55.5	46.1	ww	52.7	60.3	47.2
SD	18.7	16.9	25.6		19.8	17.2	27.2

Note. SD = Standard Deviation.

Results

The overall percentage of logically correct responses for syllogisms averaged over participants in Session 1 and 2 as well as separated for the 27 syllogisms with valid conclusions and for the 37 syllogisms without a valid conclusion (invalid syllogisms) are presented in Table 1. There were no participants with exceptionally high or low ($> |\pm 3 \text{ SDs}|$) accuracy rates in Session 1 and 2. The mean response time (RT) was 14.85 s (*SD* = 6.11 s) in Session 1 and 12.14 s (*SD* = 4.24 s) in Session 2. In follow-up questions after Session 1, 53% of all participants reported the use of at least one strategy while solving the syllogisms. For the interested reader, we included a heatmap of the response patterns in each session in Appendix D (Figure A2), including a heatmap illustrating the differences in response patterns between the sessions).

Hypothesis testing: generalized mixed model analysis

The random-effects structure of the final model included a by-participant random intercept, by-participant random slopes for *trial number, session, validity,* as well as a *validity x session* interaction including all possible correlations between the random effects. In addition, the model included a by-syllogism random intercept and a by-syllogism random slope for *session* including the correlation between them. Table 2 shows the results for the fixed effect of the final GLMM.[4]

Changes in reasoning performance within one session

In line with our prediction, syllogistic reasoning performance improved with increasing trial number. This effect was much stronger for valid than for invalid syllogisms as evident in the significant interaction between trial number and validity (see Table 2, v*alidity x trial number* interaction) as

[4] For participants, the estimated random intercept variance was 1.60, the variances of random effects of *trial number* were 0.04, of *validity* 0.71, of *session* 0.10, and 0.31 for the interaction between validity and session all for between-subject variance. For syllogisms, the estimated random intercept variance was 3.09 and 0.05 for the by-syllogisms random-slope for *session*.

Table 2. Mixed model results for the correctness of a response.

Predictors	*Odds Ratios*	*Esti-mate*	*CI*	*Std. Error*	*z*-value	*p*-value
Intercept	1.12	0.11	0.67 – 1.85	0.26	0.43	.670
Syllogism Validity (invalid)	0.66	−0.42	0.41 – 1.05	0.24	−1.74	.082
Trial Number	1.14	0.13	1.06 – 1.24	0.04	3.30	.001
Session	1.25	0.22	1.09 – 1.42	0.07	3.28	.001
Validity x Trial Number	0.88	−0.13	0.82 – 0.94	0.04	−3.71	<.001
Session x Validity	0.86	−0.15	0.73 – 1.01	0.08	−1.82	**.068**
Session x Trial Number	0.89	−0.11	0.81 – 0.99	0.05	−2.23	.025
Session x Validity x Trial Number	1.12	0.11	1.01 – 1.23	0.05	2.19	.029
Model						
Marginal R^2 / Conditional R^2				0.029 / 0.643		
Deviance				11472.404		

Note. CI = 95% confidence intervals, *df* = degrees of freedom.

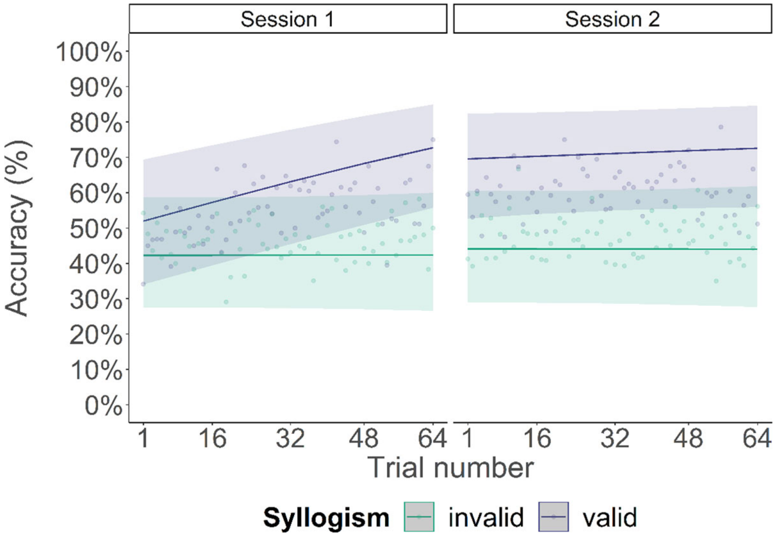


Figure 2. Estimated percentages of correct responses as a function of increasing trial number and valid as well as invalid syllogisms for Session 1 (left) and Session 2 (right).

Note. Lines represent marginal means of the mixed model. Shaded areas represent 95% confidence intervals. Data points represent means for each condition aggregated over participants.

illustrated in Figure 2. Indeed, in post-hoc analyses separated for valid and invalid syllogisms in Session 1, the slope for trial number is virtually nonexistent for invalid syllogism (regression coefficient = 0.00, p = .969) but pronounced for valid syllogisms (regression coefficient = 0.26, $p < .001$).

Retest effect

On a descriptive level, the mean rate of correct responses was only slightly higher in Session 2 than for Session 1 with a difference of 2.6% (SD = 7.6%) in correct responses. In the GLMM analysis, the likelihood to give a correct response significantly increased from Session 1 to Session 2 (see Table 2) in line with our assumption. The estimated marginal means of our model suggested a difference in reasoning performance of 5.5% (EMM_{Session1} = 52.7%, EMM_{Session1} = 58.2%). Furthermore, as predicted, a significant interaction between trial number and session was observed.

As also shown in Figure 2, the retest effect was stronger for the early trials of a session. The significant three-way interaction (also illustrated in Figure 2) indicates that this effect was stronger for valid than for invalid syllogisms. Indeed, post-hoc analyses of this effect revealed a significant retest effect only for valid (OR = 0.69, p = .002, $EMM_{\text{Session1,valid}}$ = 62.9%, $EMM_{\text{Session2,valid}}$ = 71.1%) but not for invalid (OR = 0.03, p = .435, $EMM_{\text{Session1,invalid}}$ = 42.3%, $EMM_{\text{Session2,invalid}}$ = 44.1%) syllogisms. Importantly, the slope of the covariate trend for *trial number* was only significantly different from zero for valid syllogisms in Session 1 ($p < .001$) but not for valid syllogisms in Session 2 ($p = .580$). As already mentioned, the effect of *trial number* was virtually zero and statistically non-significant for invalid syllogisms in Session 1 ($p = .989$) and likewise, it was non-significant in Session 2 ($p = .997$). Together, our analyses suggest that participants' performance improved throughout Session 1 for valid but not for invalid syllogisms. Participants were able to continue at the performance level reached at the end of Session 1 in Session 2 but did not further improve throughout this session.

Testing for individual differences in effects of time on reasoning performance

Our second analyses tested whether there exist significant individual differences between subjects in the effect magnitudes of *trial number*, *session*, as well as their interaction. Including a by-participant random slope for the *trial number x session* interaction in the final model did not substantially improve the model fit ($\chi^2 = 4.21$, $df = 6$, $p = .648$). However, excluding the by-participant random slope for the effect of *trial number* from the final model resulted in a significantly worse model fit as compared to a model without that random slope[5] ($\chi^2 = 32.04$, $df = 5$, $p < .001$). Likewise, excluding the by-participant random slope for the effect of *session* and the *session x validity* interaction resulted in a significantly worse model fit as compared to the final model including these random slopes ($\chi^2 = 52.76$, $df = 9$, $p < .001$). Thus, whereas we found no indication that participants substantially varied in the magnitude of the *trial number x session* interaction, we observed significant individual differences in the effect magnitude of *trial number* and *session*. Consequently, participants significantly differed in their improvement over time.

Associations between individual characteristics and changes in reasoning performance over time

How do the assessed cognitive ability measures and personality traits relate to the just documented individual differences in changes in reasoning performance over time? In the following, we will only briefly sum up the main results that came out of these analyses. A full report of all model fits is given in Appendix D.

[5] Excluding the random slope from the model comprises removing the corresponding correlations between that slope and the remaining random effects. Hence, more than one parameter is removed from the model for these model comparisons.

Cognitive abilities

All assessed cognitive ability measures were to some extent significantly related to participants' reasoning performance.

Raven (Appendix D, Table A3). Participant's Raven scores were strongly related to participants' reasoning performance (*OR* = 1.96, $p < .001$). The higher participants scored on the Raven, the higher their accuracies. Importantly, we found that participants' Raven scores were significantly related to both the improvement within a session and between sessions: First, we found a significant three-way interaction between the Raven, trial number, and the validity of the syllogism (*OR* = 1.10, $p = .009$; note: $p < .0125$ indicate significant results according to the Holm-Bonferroni correction, see Data Analysis section): The higher participants' Raven scores, the greater their improvement for invalid syllogisms over the time-course of an experiment (see Figure 3). Second, we found a significant three-way interaction between the Raven, session, and the validity of the (*OR* = 1.25, $p = .003$): High Raven scores were significantly associated with decreased retest effects for valid syllogisms but increased retest effects for invalid ones.

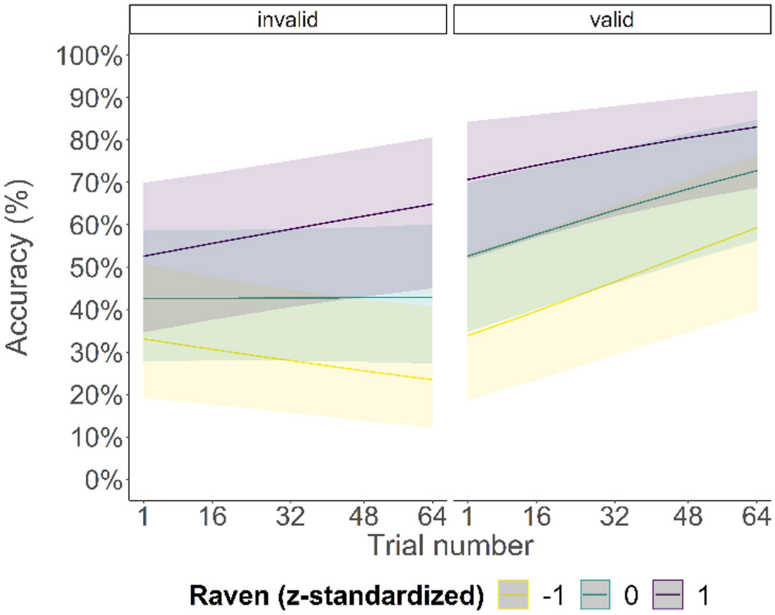


Figure 3. Estimated percentages of correct responses conditioned on Session 1 as a function of increasing trial number and standardized raven scores for invalid and valid syllogisms.

Note. Lines represent marginal means of the mixed model for average as well as 1 standard deviation below and above average Raven scores. Shaded areas represent 95% confidence intervals.

Corsi, operation span (Appendix D, Table A4), and cognitive reflection test (Table A5).

Participant's operation span ($OR = 1.38$, $p = .010$) and Cognitive Reflection Test ($OR = 1.90$, $p < .001$) scores but not participants' Corsi scores ($OR = 1.22$, $p = .119$) were significantly associated with participants' reasoning performance. Participants' operation span scores ($OR = 1.05$, $p = .235$) as well as Corsi scores ($OR = 1.05$, $p = .198$) were not significantly related to the improvement within a session. Likewise, there was no significant three-way interaction between session, the validity of the syllogism and any of the remaining cognitive ability measures: operation span ($OR = 1.17$, $p = .032$), Corsi ($OR = 1.15$, $p = .020$), and Cognitive Reflection ($OR = 1.14$, $p = .039$). Consequently, we did not find evidence that Cognitive Reflection Test, Corsi, or operation span scores were substantially related to participants' temporal changes in reasoning performance.

Personality: Need for cognition (Appendix D, Table A6) and big five (Table A7)

For the Need for Cognition model, we found neither a main effect of Need for Cognition on participants' reasoning performance nor any significant associations between temporal changes in participants' reasoning performance and their Need for Cognition scores. Thus, no results for this model are reported. With respect to the Big Five personality factors, of all factors in the model, there was only a significant main effect of Conscientiousness on participants' reasoning performance ($OR = 0.74$, $p = .018$): The lower participants' conscientiousness, the better their reasoning performance. In addition, only the factors conscientiousness ($OR = 0.89$, $p = .029$) and neuroticism ($OR = 0.93$, $p = .044$) related to participants' temporal changes in reasoning performances but failed to reach significance.

Discussion

We examined how individuals' reasoning performance changes over time, whether individuals substantially differ from one another in these temporal changes, and to what extent individuals' characteristics are associated with such interindividual differences. To this end, we investigated whether syllogistic reasoning performance increased over the course of an experiment with all 64 syllogisms as well as when being retested after a week. We measured participants' personality traits and cognitive abilities one week prior to the first syllogistic task to identify the relation between individual characteristics and the temporal improvement of reasoning performance.

Individuals reasoning performance improves within an experimental session and without feedback

For valid syllogisms, participants’ reasoning performance improved over the course of solving 64 syllogisms successively in the first assessment – in the absence of any feedback. From a test-taking perspective, various factors could have contributed to this finding. In our introduction, we elaborated on two of them: (1) the reduction of construct-irrelevant and distorting factors, and (2) the development of strategies and heuristics. We have reason to believe that specifically the development of strategies and heuristics played a major role in the observed performance gains: In follow-up questions after the first test session, at least half of the participants reported to have developed explicit strategies to solve the syllogistic reasoning tasks in Session 1. In addition, participants may have developed strategies, rules, or heuristics (such as short-cuts based on the surface features of a syllogism, e.g., Galotti et al., 1986) that could not be consciously reported. The reported use of strategies and improvements in reasoning performance is in line with other studies showing a change in students’ strategies with training (e.g., students increased their use of mental models and mental rules; Leighton, 2006) and resulting moderate improvements in performance.

Performance gains only for valid but not for invalid syllogism

In our study, participants clearly improved only for valid, but not for invalid syllogisms with increasing trial number. As 58% of the syllogisms are invalid, with NVC as the logically correct response, this finding is surprising. Importantly, we cannot attribute this observation to a general lack of potential strategies or heuristics to solve invalid syllogisms. The above-mentioned two-some rule (Galotti et al., 1986) is a good example of how participants could in principle use simple rules for invalid syllogisms to increase in speed and accuracy for the syllogisms to which it can be applied.

We offer two alternative explanations: First, strategies based on surface features such as atmosphere heuristics may suggest (false) conclusions for NVC syllogisms, leading to an improvement in valid syllogisms, but not for invalid ones. For instance, when further investigating the results patterns for each of the 64 syllogisms (see the heatmaps in Appendix D), an interesting picture emerges: For valid syllogisms, participants appear to conclude “some … are not …” (Oac and Oca) more often in Session 2 than in Session 1. This pattern was accompanied by a decline in NVC responses. While this trend resulted in more logical conclusions for some valid syllogisms (e.g., EI1, IE2, IE4), for some invalid syllogisms (e.g., EO1, OE2, OE4) responding “some … are not …” more often led to performance loss.

Second, we suggest that some individuals may be less inclined to respond NVC (earlier works already proposed such biases against the NVC response, e.g., Dickstein, 1976; Revlis, 1975; Roberts et al., 2001), which would in turn hinder the identification of logically correct strategies for invalid syllogisms. For instance, participants who assign the NVC conclusion a meaning of “giving up” (see Dames et al., 2020; Ragni et al., 2019) may either generally avoid this response option and/or put more effort into solving the problem. Revlis (1975) attributes such a NVC bias to the imbalance between valid and invalid syllogisms, and thus to an artefact of the task structure: Participants may not expect so many invalid syllogisms. Dickstein (1976) argues that participants may generally feel uncomfortable concluding that nothing follows from the given information. In addition, given the nine possible conclusions available in the task, participants may also generally underestimate the high proportion of syllogisms that require a NVC response because (e.g., when guessing) they may try to balance the use of all response options throughout the experiment. The assumption that some people show an aversion against responding NVC is also evident in a recent study showing that with feedback participants appear to become aware of the great proportion of required NVC responses in the task and consequently are able to overcome the NVC bias (Dames et al., 2020). The current finding that without feedback participants improve only for valid but not invalid syllogisms provides further support for the notion that invalid syllogisms pose particular difficulties for most participants (e.g., in form of a bias against responding NVC).

At the same time, we note that some participants may respond NVC for multiple-model problems: These participants may construct multiple models suggesting different conclusions and take this as proof that nothing follows from the premises (Bucciarelli & Johnson-Laird, 1999), missing the possibility of some common relations between the end terms of the alternative models. If that were the case, then overcoming such a bias over time could potentially explain the observed decline in NVC responses for valid syllogisms. This would explain why participants improve on average for valid but not for invalid syllogisms, as they may incorrectly apply a corresponding new reasoning strategy also to invalid problems (e.g., EO1, OE1, OE4).

Indeed, for valid syllogisms, we found that the performance difference was greater for multiple model ($M_{\text{Session2-Session1}} = +6.9\%$) than for single model problems ($M_{\text{Session2- Session1}} = +0.6\%$; $t = -3.79$, $p < .001$) supporting this notion. Yet, this result may have also been driven by the fact that reasoning performance for “easy” single model problems was, on average, already quite high in Session 1 (see the heatmap of the response patterns in Appendix D). Hence, the difference may simply stem from ceiling effects for single model problems in Session 1 and not necessarily be the result of specific underlying cognitive processes. More research is needed to disentangle when the different response biases (aversion against responding NVC vs. responding NVC for multiple model problems) can be observed and what factors bias which kind of response behavior (e.g., the nature of the task and/or interindividual differences).

Individual characteristics and differences in performance gains

Our hierarchical model analysis clearly demonstrates that participants substantially differed in the magnitude of performance gains within and between sessions: Accounting for such variations between participants significantly boosted the fit of our model to the observed data. Interestingly, beyond the influence of fluid intelligence, neither other cognitive abilities (verbal and spatial working memory capacity, Cognitive Reflection) nor personality (Need for Cognition, Big Five) measures significantly impacted participants' *temporal changes* in reasoning – at least not when introducing a stricter significance threshold to account for multiple testing. Thus, although we were able to replicate some of the typical main effects reported in previous studies (such as a negative relation between for instance neuroticism/ conscientiousness and cognitive abilities or intelligence, e.g., Carretta & Ree, 2018; Moutafi et al., 2003; Moutafi et al., 2004; Moutafi et al., 2006; Rammstedt et al., 2016), mostly individuals' fluid intelligence scores were substantially related to temporal changes in reasoning performance: Only participants who scored higher on fluid intelligence showed an improvement for invalid syllogisms over time. They may thus be some of the few participants that were able to identify useful strategies for invalid syllogisms or that were less influenced by the nature of the task (and hence potential response biases).

A note on normativism in the present study

In the present study, we assessed individuals' reasoning performance with respect to their ability to derive logically correct conclusions. We note that there is a debate on the usefulness and the pitfalls of drawing on normative standards (see Elqayam & Evans, 2011). Regardless of the discussion on normativism in reasoning research (which is not within the scope of the present study), it is worth noting that without feedback, participants' responses to valid syllogisms became increasingly normative within one session. This finding is in line with a previous study by Ball (2013) where working on multiple belief-oriented syllogisms resulted in more normative responses over time (in a non-feedback condition, this trend was stronger in a feedback condition). This raises the question of how mere reasoning practice (without feedback) and familiarity with the material can foster reasoning strategies that correspond to first-order logic.

Following Stupple and Ball (2014), one could assume that throughout the experiment, participants repeatedly balanced their intuitive inferences derived from the premises and inferences with what they believed was required or the normative standard in the present experiment (e.g., provided by the instructions or previously acquired reasoning rules). If those different types of inferences were in conflict, they may have adjusted their beliefs until they reached a justifiable trade-off between them. In line with this notion of a reflective equilibrium (see Goodman, 1983 for the general idea), Stupple and Ball (2014) proposed that during such a process, untrained individuals can align themselves with normative benchmarks without explicitly knowing that they are doing so and/or without receiving feedback. The present results converge with such a concept of informal reflective equilibrium that describes how the reasoning behavior of naïve individuals changes when given the opportunity to practice reasoning. Our findings further suggest that individuals differ in how well they succeed in such a normative alignment (or in the normative standard itself).

Individuals start off at the same level of reasoning performance, but do not further improve

Participants' reasoning performance improved when being retested. Importantly, the retest effect was stronger when comparing performance at the beginning of each test sessions: Individuals improved within the first test session and seemed to be able to start off at a similar level of performance early in the second session as reached at the end of the first one. In the second session, they then did not further improve. Such diminishing performance gains over time fit well to the power law of practice (Newell & Rosenbloom, 1981). For the syllogistic reasoning task, this result implies that if participants develop effective strategies, they are likely to do so already in the first session. They then seem to be able to use these strategies for later tests. This appears to be surprisingly fast, as a recent meta-analysis suggests that for in cognitive ability tests a plateau for performance gains through retesting is reached after the third test administration (see Scharfen et al., 2018). The authors conclude that participants typically develop such effective strategies within the first *or* second test.

The averaged retest effect in this study is nevertheless lower than a previously reported retest effect in syllogistic reasoning (Johnson-Laird & Steedman, 1978; retest effect = 10%) that was later re-analyzed by Ragni et al. (2018). However, Johnson-Laird and Steedman (1978) used a small sample ($n = 20$) restricted to Columbia University students in 1978[6]. For the mere repetition of a cognitive ability test, a recent meta-analysis reported on average an improvement of a third of a standard deviation (SD) in cognitive ability tasks (Scharfen et al., 2018). In our study, we observed an average reasoning performance of $M_{\text{Session1}} = 50.1\%$ with a $SD = 18.7\%$ in Session 1 and participants only improved 2.6% between the two sessions and not a third of a SD (6.2%). The estimated marginal means (EMMs) of our model however suggest a difference in reasoning performance of 5.5% ($EMM_{\text{Session1}} =$ 52.7%, $EMM_{\text{Session2}} = 58.2\%$) and even 8.2 % in the case of valid syllogisms (observed difference: 4.8%) which would be in line with the reported improvement in our cognitive ability measures (Scharfen et al., 2018).[7]

Our hypothesis that some capable reasoners would improve within and between sessions, was based on the assumption that individuals develop effective reasoning strategies (e.g., verbal or spatial strategies, e.g., Bacon et al., 2003) or identify rules and task-specific shortcut strategies (Roberts, 2000) in the first session and apply them in the second one. Together, our study provides support for this assumption as evident in the reported *trial number x session* interaction. In addition, around half of the participants reported the usage of rules or strategies supporting the idea that strategies were developed during the first session. At the same time, we found no relationship between personality traits and changes in reasoning performance again suggesting that the reduction of construct-irrelevant variance was not the major factor of influence.

Strength of the retest effect depends on the individual

Inspecting the strength of the retest effect again revealed substantial interindividual differences that were related to an individual's fluid

[6] In Johnson-Laird and Steedman (1978), participants received the same 64 syllogisms as used in the current study (also presented in random order for each participant). Participants had to generate their own spontaneous conclusions to each syllogism. Their two test sessions were one week apart. Hence, on the surface, the task structure was comparable to ours. One major difference however is that participants were instructed to respond as accurately and quickly as possible. In our study, there was no time pressure. Potentially, being under time pressure resulted in reasoning behavior and strategies different from our study. Although Johnson-Laird and Steedman did not report an extensive analysis of participants' RTs, they mention a reliable correlation between RTs and accuracy supporting such a notion ($r = 0.37$, $p < 0.001$). Interestingly, the authors reported that faster RTs were associated with higher accuracies. Considering that the authors did not further elaborate on whether this affected participants' improvement over time (and also considering the small sample), we refrain from further speculations on this matter.

[7] Note that the EMMs are based on the model estimates and thus differ from the observed marginal means which are based on the unmodeled. To obtain predicted values for Session 1 and Session 2 at the population-level from our model, all random effects are set to zero when estimating the EMMs. Given that individuals greatly differed in the random effect magnitude of *session*, we believe that the reported EMMs provide a suitable and more robust estimation of the retest effect than the marginal means.

intelligence. Importantly, we demonstrated that differences in the retest effect did not only depend on the individual, but, again, also on the validity of the syllogism. For valid syllogisms, retests effects were diminished for individuals with high fluid intelligence suggesting ceiling effects early in the first test session. Consequently, on the one hand, some capable participants already started with a high reasoning performance for valid syllogisms, thereby reducing the chance to improve between the sessions for those tasks. Yet, the participants that scored high on fluid intelligence demonstrated performance gains for invalid syllogisms and thus overall retest effects for invalid syllogisms. These individuals may have the resources to develop effective strategies early in the first session – even for invalid syllogisms – and apply them in the second test session.

Participants scoring low on fluid intelligence, on the other hand, performed poorer for the reasoning tasks in the first session and improved only for valid but not for invalid syllogisms: Those participants seemed to benefit relatively more from repeatedly solving the tasks and thus significantly increased in their performance over time and between test sessions for valid but not for invalid syllogisms.[8]

Implications for theories of reasoning

The current study demonstrated that individuals' reasoning performance changes over time. Such dynamic changes in reasoning performance over time are not included in most of the existing cognitive theories. Importantly, the temporal changes depend on both the type of the syllogisms and the individual. What does this imply for current models of reasoning? The results point towards differences in strategy selection or motivational factors. However, the vast majority of cognitive theories (see Khemlani & Johnson-Laird, 2012) barely considers such dynamics. Possibly, this contributes to the relatively poor performance (even on the aggregate level) of most existing cognitive theories aimed at explaining syllogistic reasoning (see Brand et al., 2019; Khemlani & Johnson-Laird, 2012; Riesterer et al., 2018; Riesterer, Brand, Dames, et al., 2020). A novel approach taken by Ackerman and Thompson (2017) may provide a promising framework to address how dynamic factors may influence reasoning. The authors introduce meta-reasoning as a meta-level processes involved in regulating and

[8] It should be noted that the study assessed the Cognitive Reflection Test and Raven at the end of Session $1+2$ after completing all 64 syllogisms. Loss of motivation and fatigue may have affected these measures. Participants performing well on the Cognitive Reflection Test and Raven at the end of a session may thus generally experience less fatigue feel less tired and more motivated. If so, however, the reported results (ceiling in retest effect as a function of participants' Raven scores) become even more striking.

monitoring reasoning (Ackerman & Thompson, 2018). Their framework is able to explain how reasoning processes change over time by taking into account meta-cognitive monitoring and control processes. It is easy to imagine how such meta-cognitive processes may in turn be influenced by individual characteristics.

In addition, our study provides evidence for substantial interindividual differences in the observed performance gains over time. Recent work demonstrated how simulations can help to determine whether it is possible for model settings to account for such individual differences between reasoners (Khemlani & Johnson-Laird, 2016). Thereby, Khemlani and Johnson-Laird (2016) showed, for instance, that the *mReasoner*[9] is capable of capturing individual differences in syllogistic reasoning between intuitive, intermediate, and deliberative reasoners. In the future, such modeling endeavors could provide a computational explanation for why some individuals improve over time when others do not.

Last, our results disagree with the assumption that syllogistic tasks measure a ‘stable’ reasoning ability. It should be noted that this finding specifically concerns the improvement in reasoning performance within the first session. At the same time, another key finding of our study is that within Session 2, participants do not further improve in their reasoning performance: a stable level seems to be reached. It is therefore possible that after some practice, stable measurements can be achieved. Further studies are needed to explore the stability of reasoning performance over larger timescales, for instance, by introducing a third condition.

Summary and conclusion

There are three main results. First, individuals improve in their syllogistic reasoning performance over the course of an experiment and they are able to start off with the same level of performance one week later when being retested. In the retest session, participants do not show further increases in their performance, resulting in a substantial retest effect especially for the first few trials of the test sessions. Second, the current study demonstrates that individuals substantially differ in the magnitude of these temporal performance gains. In addition, our results suggest that out of a variety of individual characteristics an individual’s fluid intelligence can account for some of these interindividual differences. Thus, strategies developed throughout the tasks may depend on the individual and not on “the average reasoner”.

[9] The *mReasoner* program (Khemlani & Johnson-Laird, 2013) reflects a computational implementation of the Mental Model Theory of reasoning, which proposes that during reasoning individuals construct and manipulate iconic mental representation of possibilities, i.e., mental models (e.g., Johnson-Laird, 2006).

Third, the strength of performance gains differs not only between individuals but also for the types of syllogisms: We observed performance gains over time for valid but not for invalid syllogisms. This raises the question why the validity of the syllogisms plays such a major role in dynamic changes in reasoning performance. In particular, potential influences of biases against responding NVC appear to be a promising starting point. We argued that such NVC biases are responsible for the consistently low reasoning performance for invalid syllogisms over time (even in the retest).

Together, our findings contradict common assumptions that reasoning performance (measured during one experimental session) only reflects a stable inherent ability. Rather, changes in reasoning performance differ as a function of fluid intelligence, but not as a function of the Big Five personality traits or Need for Cognition. In consequence, when testing all 64 syllogisms at once, we may not only measure an individual's ability to reason logically but also the consequences of other metacognitive processes over time, which in turn may be impacted by individual characteristics in fluid intelligence.

Acknowledgements

We thank Daniel Brand for helpful feedback on earlier versions of the manuscript.

Author note

The data and analysis scripts are publicly available on OSF: https://osf.io/x3wvf/

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the German Research Foundation (DFG) under Grant RA 1934/4-1 and RA 1934/9-1.

ORCID

Hannah Dames http://orcid.org/0000-0001-5800-9401

References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617.https://doi.org/10.1016/j.tics.2017.05.004

Ackerman, R., & Thompson, V. A. (2018). Meta-reasoning: Shedding meta-cognitive light on reasoning research. In L. J. Ball & V. A. Thompson (Eds.), *The Routledge* *international handbook of thinking and reasoning* (pp. 1–15). Routledge/Taylor & Francis Group.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bacon, A., Handley, S. J., & Newstead, S. (2003). Individual differences in strategies for syllogistic reasoning. *Thinking & Reasoning*, *9*(2), 133–168. https://doi.org/10.1080/13546780343000196

Ball, L. J. (2013). Microgenetic evidence for the beneficial effects of feedback and practice on belief bias. *Journal of Cognitive Psychology*, *25*(2), 183–191. https://doi.org/10.1080/20445911.2013.765856

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beierlein, C., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2012). *Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzskala (ASKU)*. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-292351

Beierlein, C., Kemper, C. J., Kovaleva, A., & Rammstedt, B. (2013). Short scale for measuring general self-efficacy beliefs (ASKU). *Methods, Data, Analyses*, *7*(2), 251–278. https://doi.org/10.12758/mda.2013.014

Bender, A., Beller, S., & Klauer, K. C. (2016). Lady Liberty and Godfather Death as candidates for linguistic relativity? Scrutinizing the gender congruency effect on personified allegories with explicit and implicit measures. *Quarterly Journal of Experimental Psychology*, *69*(1), 48–64. https://doi.org/10.1080/17470218.2015.1021701

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F., & Schwarz, N. (1994). Need for cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben. *Zeitschrift Für Sozialpsychologie*, *25*, 147–154.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, *58*(3), 382–398. https://doi.org/10.1177/0013164498058003002

Brand, D., Riesterer, N., & Ragni, M. (2019). On the matter of aggregate models for syllogistic reasoning: A transitive set-based account for predicting the population. In T. Stewart (Chair), *Proceedings of the 17th International Conference on Cognitive Modeling*.

Brase, G. L., Osborne, E. R., & Brandner, J. L. (2019). General and specific personality traits as predictors of domain-specific and general conditional reasoning. *Personality and Individual Differences*, *137*, 157–164. https://doi.org/10.1016/j.paid.2018.08.017

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*(3), 247–303. https://doi.org/10.1207/s15516709cog2303_1

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Carretta, T. R., & Ree, M. J. (2018). The relations between cognitive ability and personality: Convergent results across measures. *International Journal of Selection and Assessment*, *26*(2–4), 133–144. https://doi.org/10.1111/ijsa.12224

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*(2), 191–258. https://doi.org/10.1006/cogp.1998.0696

Copeland, D., & Radvansky, G. (2004). Working memory and syllogistic reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *57*(8), 1437–1457. https://doi.org/10.1080/02724980343000846

Corsi, P. M. (1973). *Human memory and the medial temporal region of the brain* [Doctoral dissertation]. ProQuest Information & Learning.

Dames, H., Schiebel, C., & Ragni, M. (2020). The role of feedback and post-error adaptations in reasoning. In *Proceedings of the 42th Annual Conference of the Cognitive Science Society. Cognitive Science Society* (pp. 3275–3281). Cognitive Science Society.

Dickstein, L. S. (1976). Differential difficulty of categorical syllogisms. *Bulletin of the Psychonomic Society*, *8*(4), 330–332. https://doi.org/10.3758/BF03335156

Elqayam, S., & Evans, J. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*(5), 233–248. https://doi.org/10.1017/S0140525X1100001X

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

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*(2), 336–353. https://doi.org/10.1037/1528-3542.7.2.336

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Freund, P. A., & Holling, H. (2011). Who wants to take an intelligence test? Personality and achievement motivation in the context of ability testing. *Personality and Individual Differences*, *50*(5), 723–728. https://doi.org/10.1016/j.paid.2010.12.025

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

Galotti, K. M., Baron, J., & Sabini, J. P. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology. General*, *115*(1), 16–25.https://doi.org/10.1037/0096-3445.115.1.16

Gerlitz, J. Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. *DIW Research Notes*, *4*. https://www.diw.de/documents/dokumentenarchiv/17/43490/rn4.pdf

Goodman, N. (1983). *Fact, fiction, and forecast*. Harvard University Press.

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology*, *87*(2), 243–254. https://doi.org/10.1037/0021-9010.87.2.243

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence*, *48*, 1–14.https://doi.org/10.1016/j.intell.2014.10.005

Johnson-Laird, P. N. (2006). *How we reason*. Oxford University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Lawrence Erlbaum Associates, Inc.

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive Psychology*, *10*(1), 64–99. https://doi.org/10.1016/0010-0285(78)90019-1

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science*, *4*(1), 71–115. https://doi.org/10.1016/S0364-0213(81)80005-5

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69.https://doi.org/10.1037/a0028347

Khemlani, S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, *4*(1), 4–20. https://doi.org/10.1080/19462166.2012.674060

Khemlani, S., & Johnson-Laird, P. N. (2016). How people differ in syllogistic reasoning. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2165–2170). Cognitive Science Society.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–457.https://doi.org/10.1037/a0026841

Khemlani, S., & Johnson-Laird, P. N. (2016). How people differ in syllogistic reasoning. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Ed.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Leighton, J. P. (2006). Teaching and assessing deductive reasoning skills. *The Journal of Experimental Education*, *74*(2), 107–136. https://doi.org/10.3200/JEXE.74.2.107-136

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *The Journal of Applied Psychology*, *92*(6), 1672–1682.https://doi.org/10.1037/0021-9010.92.6.1672

Lüdecke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772. https://doi.org/10.21105/joss.00772

Matton, N., Vautier, S., & Raufaste, É. (2009). Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence*, *37*(4), 412–421. https://doi.org/10.1016/j.intell.2009.03.011

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin*, *27*(3), 272–277.

Moutafi, J., Furnham, A., & Crump, J. (2003). Demographic and personality predictors of intelligence: A study using the Neo Personality Inventory and the Myers–Briggs Type Indicator. *European Journal of Personality*, *17*(1), 79–94. https://doi.org/10.1002/per.471

Moutafi, J., Furnham, A., & Paltiel, L. (2004). Why is Conscientiousness negatively correlated with intelligence? *Personality and Individual Differences*, *37*(5), 1013–1022. https://doi.org/10.1016/j.paid.2003.11.010

Moutafi, J., Furnham, A., & Tsaousis, I. (2006). Is the relationship between intelligence and trait Neuroticism mediated by test anxiety? *Personality and Individual Differences*, *40*(3), 587–597. https://doi.org/10.1016/j.paid.2005.08.004

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination *R2* and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive Skills and Their Acquisition*, *1*, 1–55.

Newstead, S. E., Stephen, E., Handley, S. J., Harley, C., Wright, H., & Farrelly, D. (2004). Individual differences in deductive reasoning. *The Quarterly Journal of*

Experimental Psychology. A, Human Experimental Psychology, *57*(1), 33–60.https://doi.org/10.1080/02724980343000116

Ng, E., & Lee, K. (2015). Effects of trait test anxiety and state anxiety on children's working memory task performance. *Learning and Individual Differences*, *40*, 141–148. https://doi.org/10.1016/j.lindif.2015.04.007

Ragni, M., Dames, H., Brand, D., & Riesterer, N. (2019). When does a reasoner respond: Nothing follows? In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 2640–2646). Cognitive Science Society. * contributed equally

Ragni, M., Riesterer, N., Khemlani, S., & Johnson-Laird, P. N. (2018). Individuals become more logical without feedback. In T. Rogers, M. Rau, J. Zhu, & C. Kalish (Ed.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1584–1589).Cognitive Science Society.

Rammstedt, B., Danner, D., & Martin, S. (2016). The association between personality and cognitive ability: Going beyond simple effects. *Journal of Research in Personality*, *62*, 39–44. https://doi.org/10.1016/j.jrp.2016.03.005

Reeve, C. L., Heggestad, E. D., & Lievens, F. (2009). Modeling the impact of test anxiety and test familiarity on the criterion-related validity of cognitive ability tests. *Intelligence*, *37*(1), 34–41. https://doi.org/10.1016/j.intell.2008.05.003

Revlis, R. (1975). Syllogistic reasoning: Logical decisions from a complex data base. In R. Falmagne (Ed.), *Reasoning: Representation and process* (pp. 93–133). Erlbaum.

Riesterer, N., Brand, D., Dames, H., & Ragni, M. (2020). Modeling human syllogistic reasoning: The role of "no valid conclusion". *Topics in Cognitive Science*, *12*(1), 446–459.

Riesterer, N., Brand, D., & Ragni, M. (2018). The predictive power of heuristic portfolios in human syllogistic reasoning. In F. Trollmann & A.-Y. Turhan (Eds.), *Proceedings of the 41st German Conference on AI* (pp. 415–421). Springer 11:44 Uhr.

Riesterer, N., Brand, D., & Ragni, M. (2020). Predictive modeling of individual human cognition: Upper bounds and a new perspective on performance. *Topics in Cognitive Science*, *12*, 960–974.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. MIT Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1679

Roberts, M. J., & Newton, E J. (2003). Individual differences in the development of reasoning strategies. In D. K. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment and decision making* (pp. 23–43). J. Wiley.

Roberts, M. J., Newstead, S. E., & Griggs, R. A. (2001). Quantifier interpretation and syllogistic reasoning. *Thinking & Reasoning*, *7*(2), 173–204. https://doi.org/10.1080/13546780143000008

Roberts, M. J. (2000). Individual differences in reasoning strategies: A problem to solve or an opportunity to seize. In W. Schaeken, G. de Vooght, A. Vandierendonck, & G. d'Ydewalle (Eds.), *Deductive reasoning and strategies* (pp. 23–48). L. Erlbaum Associates.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.https://doi.org/10.1016/j.tics.2010.09.003

Schaeken, W., de Vooght, G., Vandierendonck, A., & d'Ydewalle, G. (Eds.). (2000). *Deductive reasoning and strategies*. L. Erlbaum Associates.

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence*, *67*, 44–66. https://doi.org/10.1016/j.intell.2018.01.003

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2017). *Afex: Analysis of factorial experiments*. R package version 0.18-0. https://CRAN.R-project.org/package=afex.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Stone, J. M., & Towse, J. N. (2015). A working memory test battery: Java-based collection of seven working memory tasks. *Journal of Open Research Software*, *3*(2), 92. https://doi.org/10.5334/jors.br

Störring, G. (1908). Experimentelle Untersuchungen über einfache Schlußprozesse. *Archiv Für Die Gesamte Psychologie*, (11), 1–27.

Stupple, E. J. N., & Ball, L. J. (2014). The intersection between descriptivism and meliorism in reasoning research: Further proposals in support of 'soft normativism. *Frontiers in Psychology*, *5*, 1269. https://doi.org/10.3389/fpsyg.2014.01269

Süß, H. M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence*, *30*(3), 261–288. https://doi.org/10.1016/S0160-2896(01)00100-3

Svedholm-Häkkinen, A. M. (2015). Highly reflective reasoners show no signs of belief inhibition. *Acta Psychologica*, *154*, 69–76.https://doi.org/10.1016/j.actpsy.2014.11.008

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de L'Academie Canadienne de Psychiatrie de L'enfant et de L'adolescent*, *19*(3), 227–229.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, *18*(4), 451–460. https://doi.org/10.1037/h0060520

Appendix A

Syllogism tasks

Table A1. Items used in the syllogistic reasoning task.

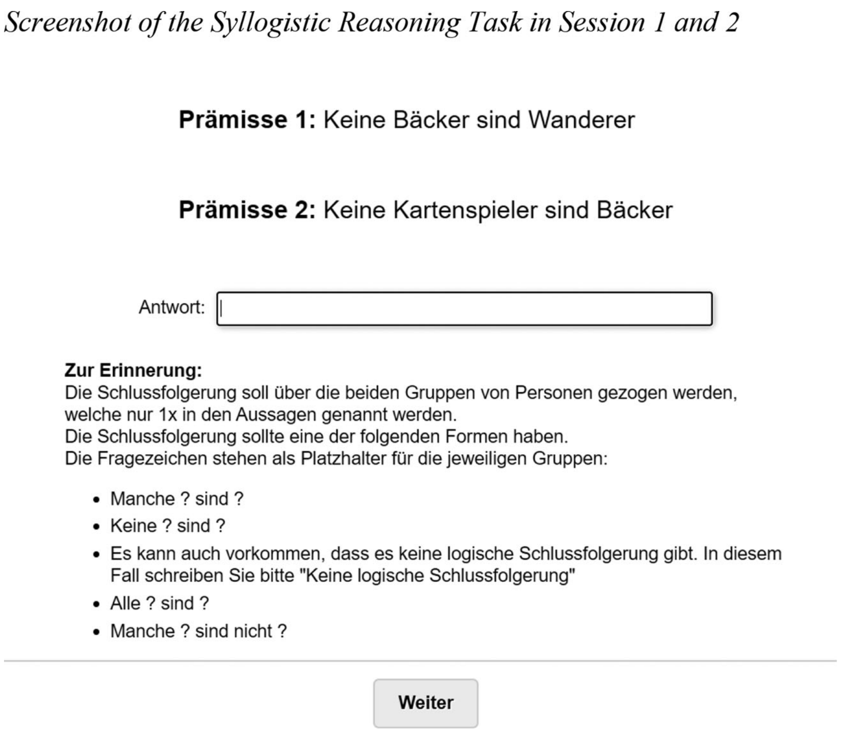
Occupations	Sports	Hobbies
Anwälte	Jogger	Gärtner
Ärzte	Kraftsportler	Computerspieler
Bäcker	Schwimmer	Heimwerker
Beamte	Taucher	Tüftler
Chemiker	Radfahrer	Sänger
Chirurgen	Wanderer	Camper
Dachdecker	Kletterer	Fotografen
Elektroniker	Tänzer	Musiker
Erzieher	Reiter	Künstler
Friseure	Turner	Maler
Informatiker	Angler	Zeichner
Ingenieure	Jäger	Bücherleser

(*continued*)

Table A1. Continued.

Occupations	Sports	Hobbies
Journalisten	Kampfsportler	Schauspieler
Juristen	Golfer	Kartenspieler
Kellner	Boxer	Töpfer
Klempner	Segler	Trommler
Köche	Fußballer	Pianisten
Lehrer	Handballspieler	Gitarristen
Maurer	Tennisspieler	Geiger
Mechaniker	Basketballspieler	Flötenspieler
Piloten	Volleyballspieler	Schlagzeuger
Politiker	Handballspieler	Trompeter
Polizisten	Eishockeyspieler	Feinschmecker
Optiker	Wanderer	Bassisten
Richter	Bergsteiger	Zeitungsleser
Sanitäter	Surfer	Sammler
Schneider	Sprinter	Kartenspieler
Schreiner	Skifahrer	Zauberer
Soldaten	Rennfahrer	Dichter
Techniker	Schlittschuhläufer	Autoren
Verkäufer	Badminton-Spieler	Erzähler

Appendix B



Screenshot of the Syllogistic Reasoning Task in Session 1 and 2

Prämisse 1: Keine Bäcker sind Wanderer

Prämisse 2: Keine Kartenspieler sind Bäcker

Antwort:

Zur Erinnerung:
Die Schlussfolgerung soll über die beiden Gruppen von Personen gezogen werden, welche nur 1x in den Aussagen genannt werden.
Die Schlussfolgerung sollte eine der folgenden Formen haben.
Die Fragezeichen stehen als Platzhalter für die jeweiligen Gruppen:

- Manche ? sind ?
- Keine ? sind ?
- Es kann auch vorkommen, dass es keine logische Schlussfolgerung gibt. In diesem Fall schreiben Sie bitte "Keine logische Schlussfolgerung"
- Alle ? sind ?
- Manche ? sind nicht ?

Figure A1. Screenshot of the syllogistic reasoning task in Session 1 and 2. *Note.* English translation for the screenshot: “

Premise 1: No backers are hikers
Premise 2: No card players are bakers
Response: [___]
Reminder:
You should draw a conclusion about the two groups of people that are named only once in the statements.
The conclusion should take one of the following forms.
The question marks represent placeholders for the corresponding groups:

- Some ? are ?
- No ? are ?
- There may be cases in which no valid conclusion can be inferred. In such a case "no conclusion is possible" can be answered
- All ? are ?
- Some ? are not ?

Appendix C

To determine the most appropriate random effects structure for our model, we first fitted our model with a maximal random effects structure justified by the design including: by-participant and by-syllogism random intercepts and by-participant random slopes for _session_, _validity_, _trial number_, and all interactions and correlations for participants as well as by-syllogism random slopes for _session_, _trial number_, and their interactions and correlations (R-formula: _correctness ~ validity*trial-number*session + (trial-number *session|syllogism) + (validity* trial-number *session|participant)_). This model did not successfully converge. Step-by-step we then reduced the random effect structure (starting with excluding correlations for higher-order interactional random effects, then higher-order interactions, and so on). The following model was determined to be the "full" model with no convergence issues:

Model$_{\text{full}}$: _correctness ~ validity*trial-number*session + (trial-number *session|syllogism) + (trial-number + validity*session|participant)_

In a next step the most appropriate random effect structure was estimated. This model selection procedure was based on Bender et al. (2016) work. We compared this model with a "null" model with only by-participant and by-syllogism random intercepts, thus without any random slopes. If the null model had explained the data as well as the full model, we would have accepted the null model as the final model (for the analysis of the fixed effects). However, the "null" model fitted the data significantly worse than the "full" model (see Table A2) suggesting that our final model required additional random effects. To investigate which random slopes were required in our final model, we inspected the estimated variance for each random slope in the full model. Using a stepwise approach, we selected the random slopes with the largest variance in the full model (e.g., by-participant random slope for the factor validity in a first step). We then fitted "reduced" models including these random slopes and contrasted them with the full model. If the reduced model fit the data as well as the full model, we would keep the reduced model as a final model. All models fitted that way are reported in Table A2.

The reduced models were specified as follows (R connotation):

$M_{minimal}$: _correctness ~ validity * trial number * session + (1 | syllogism) + (1 |participant)_

$M_{reduced1}$ (add by-participant random slope for validity): _correctness ~ validity * trial number * session + (1 | syllogism) + (validity |participant)_

$M_{reduced2}$ (add by-participant random slope for session): _correctness ~ validity * trial number * session + (1 | syllogism) + (validity + session | participant)_

$M_{reduced3}$ (add by-participant random slope for _validity x session_ interaction): _correctness ~ validity * trial number * session + (1 | syllogism) + (validity*session | participant)_

$M_{reduced4}$ (add by-syllogism random slope for session): _correctness ~ validity * trial number * session + (session | syllogism) + (validity*session | participant)_

$M_{reduced5}$ (add by-participant random slope for trial number): _correctness ~ validity * trial number * session + (session | syllogism) +_
(trial number + validity*session | participant)

Table A2. Mixed models fitted to estimate the most appropriate random effect structure.

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	Δx^2	Δdf	*p*
M_{full}	33	11536	11782	−5735.1	11470			
$M_{minimal}$	10	12313	12388	−6146	12293	822.71	23	<.001
$M_{reduced1}$	12	11583	11672	−5779.5	11559	88.697	21	<.001
$M_{reduced2}$	15	11583	11694	−5776.3	11553	82.344	18	<.001
$M_{reduced3}$	19	11546	11687	−5753.8	11508	37.428	14	<.001
$M_{reduced4}$	21	11546	11703	−5752.2	11504	34.174	12	<.001
$M_{reduced5}$ & M_{final}	26	11524	11718	−5736.2	11472	2.137	7	.952

Note. All models are compared to M_{full}.

Appendix D

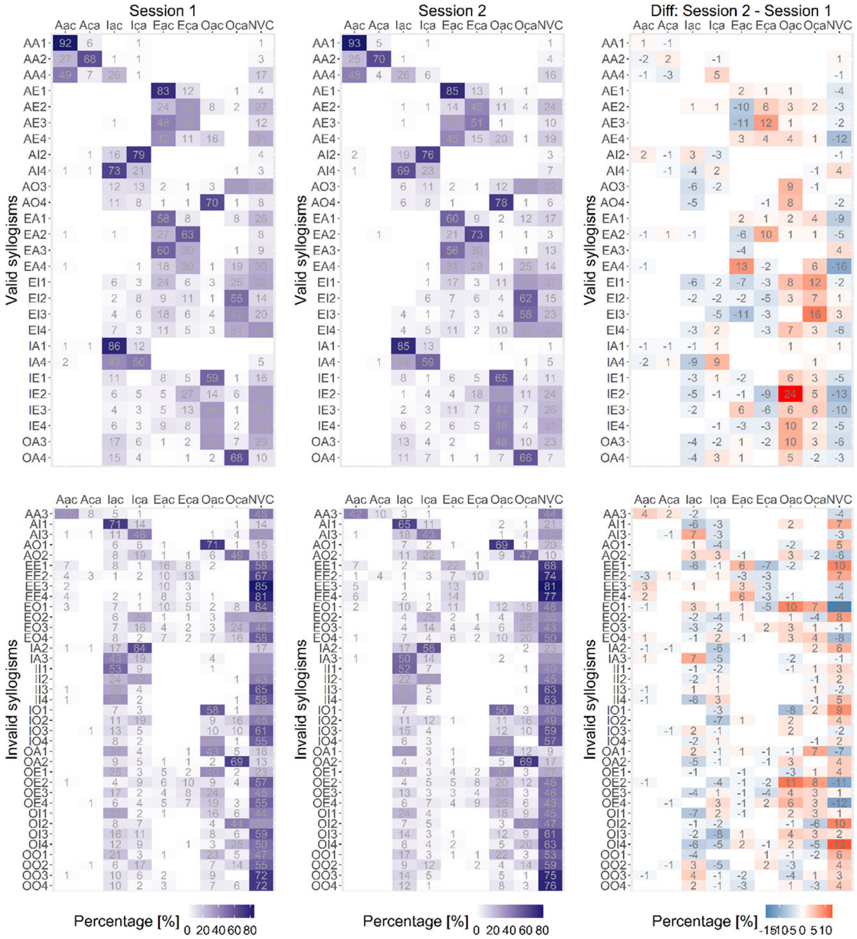


Figure A2. Heatmap of the response patterns in Session 1 and Session 2.

Note. For nine conclusions (columns), cell colors reflect the percentages of responses for the 64 syllogisms (rows) in Session 1 (left), Session 2 (middle), and between the two Sessions (right; blue/red colors reflect fewer/more responses in Session 2 than in Session 1). The upper 27 rows denote syllogisms with a valid conclusion and the lower 37 denote invalid syllogisms. Syllogisms labels: A, I, E, and O represent the quantifiers “All”, “Some”, “No”, and “Some … not”, numbers reflect the syllogistic figure, i.e., the order of terms in the premises; Aac = All of the A are C, Iac = Some of the A are C, Eac = None of the A is a C, Oac = Some of the A are not C, and NVC = no valid conclusion.

Appendix E

Table A3. Mixed model results for the influence of raven on the correctness of a response.

Predictors	*Odds Ratios*	*CI*	*Std. Error*	*Statistic*	*p*
Intercept	1.13	0.69 – 1.85	0.25	0.50	.615
Syllogism Validity (invalid)	0.66	0.41 – 1.06	0.24	−1.73	.083
Trial Number	1.14	1.05 – 1.23	0.04	3.30	**.001**
Retest (Session 2)	1.25	1.10 – 1.43	0.07	3.31	**.001**
Raven	1.96	1.55 – 2.46	0.12	5.69	**<.001**
Syllogism Validity x Trial Number	0.88	0.82 – 0.94	0.04	−3.58	**<.001**
Retest x Syllogism Validity	0.87	0.74 – 1.01	0.08	−1.81	.070
Retest x Trial Number	0.90	0.81 – 0.99	0.05	−2.14	**.033**
Raven x Syllogism Validity	0.98	0.81 – 1.18	0.09	−0.21	.835
Raven x Trial Number	1.05	0.97 – 1.14	0.04	1.16	.248
Retest x Raven	1.02	0.90 – 1.15	0.06	0.26	.791
Retest x Syllogism Validity x Trial Number	1.11	1.01 – 1.23	0.05	2.15	**.032**
Raven x Syllogism Validity x Trial Number	1.10	1.02 – 1.19	0.04	2.61	**.009**
Retest x Syllogism Validity x Raven	1.25	1.08 – 1.44	0.07	2.96	**.003**
Retest x Trial Number x Raven	1.05	0.94 – 1.16	0.05	0.85	.393
Retest x Syllogism Validity x Trial Number x Raven	0.96	0.86 – 1.06	0.05	−0.87	.385
Observations		12800			
Marginal R^2 / Conditional R^2		0.082 / 0.644			

Note. CI = 95% confidence intervals, *df* = degrees of freedom. *p*-values printed in bold indicate significant effects when compared to the corresponding alpha levels adjusted for multiple testing using the Holm-Bonferroni Method (adjusted alpha levels for cognitive ability measures: .0125, .0167, .025, .05 and for personality measures: .01, .0125, .0167, .025, .05).

Table A4. Mixed model results for the influence of corsi and operation span on the correctness of a response.

Predictors	*Estimate*	*CI*	*Std. Error*	*Statistic*	*p*
Intercept	1.12	0.68 – 1.85	0.29	0.45	.656
Syllogism Validity (invalid)	0.66	0.41 – 1.06	0.16	−1.72	.085
Trial Number	1.14	1.06 – 1.23	0.05	3.33	**.001**
Retest (Session 2)	1.25	1.10 – 1.42	0.08	3.42	**.001**
Operation Span	1.38	1.08 – 1.77	0.17	2.58	**.010**
Corsi	1.22	0.95 – 1.56	0.15	1.56	.119
Syllogism Validity x Trial Number	0.88	0.82 – 0.94	0.03	−3.68	**<.001**
Retest x Syllogism Validity	0.86	0.73 – 1.01	0.07	−1.90	.058
Retest x Trial Number	0.89	0.81 – 0.99	0.04	−2.24	**.025**
Operation Span x Syllogism Validity	0.90	0.75 – 1.07	0.08	−1.17	.242
Operation Span x Trial Number	1.05	0.97 – 1.13	0.04	1.19	.235
Retest x Operation Span	1.02	0.91 – 1.15	0.06	0.37	.708
Corsi x Syllogism Validity	1.12	0.94 – 1.34	0.10	1.24	.216
Corsi x Trial Number	1.05	0.97 – 1.14	0.04	1.29	.198
Retest x Corsi	1.15	1.02 – 1.28	0.07	2.33	.020
Retest x Syllogism Validity x Trial Number	1.12	1.01 – 1.23	0.06	2.19	.028
Operation Span x Syllogism Validity x Trial Number	1.04	0.97 – 1.11	0.04	1.08	.281
Retest x Syllogism Validity x Operation Span	1.17	1.01 – 1.36	0.09	2.14	.032
Retest x Trial Number x Operation Span	1.03	0.93 – 1.14	0.05	0.52	.603
Corsi x Syllogism Validity x Trial Number	0.99	0.92 – 1.06	0.04	−0.24	.812
Retest x Syllogism Validity x Corsi	0.94	0.81 – 1.09	0.07	−0.86	.390
Retest x Trial Number x Corsi	0.96	0.87 – 1.06	0.05	−0.79	.429
Retest x Syllogism Validity x Trial Number x Operation Span	0.98	0.88 – 1.08	0.05	−0.44	.659
Retest x Syllogism Validity x Trial Number x Corsi	1.05	0.95 – 1.16	0.05	0.94	.347
Observations		12800			
Marginal R^2 / Conditional R^2		0.055 / 0.643			

Note. CI = 95% confidence intervals, *df* = degrees of freedom. *p*-values printed in bold indicate significant effects when compared to the corresponding alpha levels adjusted for multiple testing using the Holm-Bonferroni Method (adjusted alpha levels for cognitive ability measures: .0125, .0167, .025, .05 and for personality measures: .01, .0125, .0167, .025, .05).

Table A5. Mixed model results for the influence of cognitive reflection test (CRT)on the correctness of a response.

Predictors	*Estimate*	*CI*	*Std. Error*	*Statistic*	*p*
Intercept	1.12	0.69 – 1.83	0.28	0.45	.650
Syllogism Validity (invalid)	0.66	0.41 – 1.06	0.16	−1.74	.083
Trial Number	1.14	1.06 – 1.23	0.05	3.33	**.001**
Retest (Session 2)	1.25	1.10 – 1.43	0.08	3.42	**.001**
CRT	1.90	1.51 – 2.40	0.22	5.46	**<.001**
Syllogism Validity x Trial Number	0.88	0.82 – 0.94	0.03	−3.67	**<.001**
Retest x Syllogism Validity	0.86	0.74 – 1.01	0.07	−1.82	.070
Retest x Trial Number	0.89	0.81 – 0.99	0.04	−2.23	.026
CRT x Syllogism Validity	1.04	0.86 – 1.25	0.10	0.40	.689
CRT x Trial Number	1.05	0.97 – 1.14	0.04	1.17	.242
Retest x CRT	1.14	1.01 – 1.28	0.07	2.06	.039
Retest x Syllogism Validity x Trial Number	1.12	1.01 – 1.23	0.06	2.21	**.027**
CRT x Syllogism Validity x Trial Number	1.06	0.98 – 1.13	0.04	1.50	.133
Retest x Syllogism Validity x CRT	1.12	0.96 – 1.30	0.09	1.49	.136
Retest x Trial Number x CRT	1.06	0.96 – 1.17	0.05	1.11	.267
Retest x Syllogism Validity x Trial Number x CRT	1.00	0.90 – 1.10	0.05	−0.06	.954
Observations	12800				
Marginal R^2 / Conditional R^2	0.086 / 0.643				

Note. CI = 95% confidence intervals, *df* = degrees of freedom. *p*-values printed in bold indicate significant effects when compared to the corresponding alpha levels adjusted for multiple testing using the Holm-Bonferroni Method (adjusted alpha levels for cognitive ability measures: .0125, .0167, .025, .05 and for personality measures: .01, .0125, .0167, .025, .05)..

Table A6. Mixed model results for the influence of need for cognition (NFC) on the correctness of a response.

Predictors	*Estimate*	*CI*	*Std. Error*	*Statistic*	*p*
Intercept	1.12	0.67 – 1.85	0.29	0.43	.670
Syllogism Validity (invalid)	0.66	0.41 – 1.05	0.16	−1.74	.082
Trial Number	1.14	1.06 – 1.24	0.05	3.32	**.001**
Retest (Session 2)	1.25	1.09 – 1.42	0.08	3.29	**.001**
NFC	1.10	0.85 – 1.43	0.15	0.72	.470
Syllogism Validity x Trial Number	0.88	0.82 – 0.94	0.03	−3.70	**<.001**
Retest x Syllogism Validity	0.86	0.73 – 1.01	0.07	−1.82	.069
Retest x Trial Number	0.89	0.81 – 0.99	0.04	−2.24	**.025**
NFC x Syllogism Validity	0.99	0.83 – 1.19	0.09	−0.10	.919
NFC x Trial Number	1.01	0.93 – 1.09	0.04	0.18	.861
Retest x NFC	0.97	0.86 – 1.09	0.06	−0.56	.578
Retest x Syllogism Validity x Trial Number	1.12	1.01 – 1.23	0.06	2.18	**.030**
NFC x Syllogism Validity x Trial Number	1.03	0.96 – 1.11	0.04	0.86	.391
Retest x Syllogism Validity x NFC	1.08	0.93 – 1.25	0.08	0.98	.326
Retest x Trial Number x NFC	0.99	0.90 – 1.09	0.05	−0.23	.818
Retest x Syllogism Validity x Trial Number x NFC	0.95	0.86 – 1.05	0.05	−0.97	.332
Observations	12800				
Marginal R^2 / Conditional R^2	0.030 / 0.643				

Note. CI = 95% confidence intervals, *df* = degrees of freedom. *p*-values printed in bold indicate significant effects when compared to the corresponding alpha levels adjusted for multiple testing using the Holm-Bonferroni Method (adjusted alpha levels for cognitive ability measures: .0125, .0167, .025, .05 and for personality measures: .01, .0125, .0167, .025, .05).

Table A7. Mixed model results for the influence of the big five factors on the correctness of a response.

Predictors	*Odds Ratios*	*CI*	*Std. Error*	*Statistic*	*p*
Intercept	1.10	0.67 – 1.83	0.26	0.39	.699
Syllogism Validity (invalid)	0.65	0.41 – 1.05	0.24	−1.77	.076
Trial Number	1.14	1.05 – 1.23	0.04	3.27	**.001**
Session	1.25	1.10 – 1.42	0.07	3.32	**.001**
Openness	1.19	0.92 – 1.55	0.13	1.31	.189
Extraversion	0.94	0.72 – 1.24	0.14	−0.43	.670
Conscientiousness	0.74	0.58 – 0.95	0.13	−2.37	**.018**
Neuroticism	0.85	0.65 – 1.10	0.13	−1.25	.210
Validity x Trial Number	0.87	0.81 – 0.94	0.04	−3.80	**<.001**
Validity x Session	0.86	0.74 – 1.01	0.08	−1.80	.071
Trial Number x Session	0.90	0.82 – 1.00	0.05	−2.03	**.042**
Validity x Openness	1.06	0.88 – 1.28	0.10	0.64	.525
Trial Number x Openness	0.99	0.91 – 1.08	0.04	−0.15	.883
Session x Openness	0.99	0.87 – 1.12	0.06	−0.21	.836
Validity x Extraversion	0.87	0.72 – 1.06	0.10	−1.41	.159
Trial Number x Extraversion	1.01	0.93 – 1.10	0.04	0.32	.747
Session x Extraversion	1.03	0.90 – 1.16	0.06	0.39	.694
Validity x Conscientiousness	0.90	0.76 – 1.08	0.09	−1.15	.251
Trial Number x Conscientiousness	1.03	0.95 – 1.11	0.04	0.72	.470
Session x Conscientiousness	1.00	0.89 – 1.13	0.06	0.06	.949
Validity x Neuroticism	0.90	0.75 – 1.08	0.09	−1.10	.272
Trial Number x Neuroticism	1.05	0.96 – 1.14	0.04	1.07	.284
Session x Neuroticism	1.06	0.94 – 1.20	0.06	0.96	.335
Validity x Trial Number x Session	1.12	1.01 – 1.23	0.05	2.18	**.029**
Validity x Trial Number x Openness	1.04	0.96 – 1.12	0.04	0.98	.328
Validity x Session x Openness	0.89	0.76 – 1.04	0.08	−1.50	.134
Trial Number x Session x Openness	0.97	0.87 – 1.08	0.05	−0.57	.567
Validity x Trial Number x Extraversion	0.97	0.90 – 1.05	0.04	−0.72	.473
Validity x Session x Extraversion	0.96	0.82 – 1.12	0.08	−0.53	.596
Trial Number x Session x Extraversion	0.94	0.85 – 1.05	0.05	−1.08	.280
Validity x Trial Number x Conscientiousness	0.94	0.87 – 1.01	0.04	−1.81	.070
Validity x Session x Conscientiousness	1.00	0.86 – 1.15	0.07	−0.05	.963
Trial Number x Session x Conscientiousness	0.89	0.81 – 0.99	0.05	−2.19	.029
Validity x Trial Number x Neuroticism	0.93	0.86 – 1.00	0.04	−2.02	.044
Validity x Session x Neuroticism	0.97	0.84 – 1.14	0.08	−0.32	.745
Trial Number x Session x Neuroticism	1.04	0.94 – 1.15	0.05	0.73	.466
Validity x Trial Number x Session x Openness	0.92	0.82 – 1.02	0.05	−1.66	.097
Validity x Trial Number x Session x Extraversion	1.04	0.93 – 1.15	0.05	0.65	.517
Validity x Trial Number x Session x Conscientiousness	1.09	0.99 – 1.21	0.05	1.73	.083
Validity x Trial Number x Session x Neuroticism	1.03	0.92 – 1.14	0.05	0.47	.636
Model					
Marginal R^2 / Conditional R^2			0.052 / 0.645		
Observations			12800		

Note. CI = 95% confidence intervals, *df* = degrees of freedom. *p*-values printed in bold indicate significant effects when compared to the corresponding alpha levels adjusted for multiple testing using the Holm-Bonferroni Method (adjusted alpha levels for cognitive ability measures: .0125, .0167, .025, .05 and for personality measures: .01, .0125, .0167, .025, .05).