Reasoning ability measures, omnipresent, yet not fully understood. A closer look at the learning hypothesis.

Inauguraldissertation der Philosophisch-humanwissenschaftlichen Fakultät der Universität Bern zur Erlangung der Doktorwürde vorgelegt von:

Helene M. von Gugelberg

Maienfeld, September 2024

Table of Contents

Abstract

Human intelligence and therefore measures of reasoning ability have been state of the art for predicting potential success for an individual. Yet much of individual differences in the test taking process itself are still unclear. Only when we understand what influences test taking behavior and its outcome, can we aim for test fairness across individuals and cultures. The three studies presented this dissertation took a closer look at the item-position effect under the premise of the learning hypothesis. The item-position effect captures the often discovered increasing (true) item variance within a reasoning ability measure with homogenous items in addition to a latent variable depicting reasoning. The learning hypothesis postulates, that this increase in variance is due to individual differences in the ability to learn the underlying rules of items during the test taking process. By conducting three empirical studies, cognitive, behavioral, and methodological factors contributing to this phenomenon are investigated. Study 1 (von Gugelberg et al., 2021) linked the item-position effect to proactive mechanism of control. Study 2 (von Gugelberg & Troche, *in preparation*) found a shift towards more effective strategy use related to a more pronounced item-position effect and in study 3 (von Gugelberg et al. 2025) the item-position effect was disrupted by an experimental manipulation of the underlying rules within a reasoning test. The presented dissertation made the investigation of the item-position effect more accessible by creating openly availably tests for reasoning ability and detailed explanations of the fixed-links approach based in R, a frequently used freeware for statistical analysis. Further, presented results support the learning hypothesis, albeit not unambiguously. Alternative explanations and future study designs are provided in detail. Additionally, a broader definition of the item-position effect is proposed, based on the current state of evidence. Future research should investigate whether individual differences in adaptive behavior during test taking describes the phenomenon underlying the item-position effect more accurately.

Introduction

Intelligence, or as we nowadays must specify, human intelligence is an imperfect (unresolved) puzzle that captivates many researchers. It long has been used as the holy grail for predicting the potential success of an individual in school (Deary et al., 2007; Roth et al., 2015), chances to enter a university (Davey et al., 2007), and later job performance (Schmidt & Hunter, 1998). Despite the broad usage of the construct of human intelligence, there are still ongoing debates about the underlying structure and what measures are up to the task of providing unbiased information.

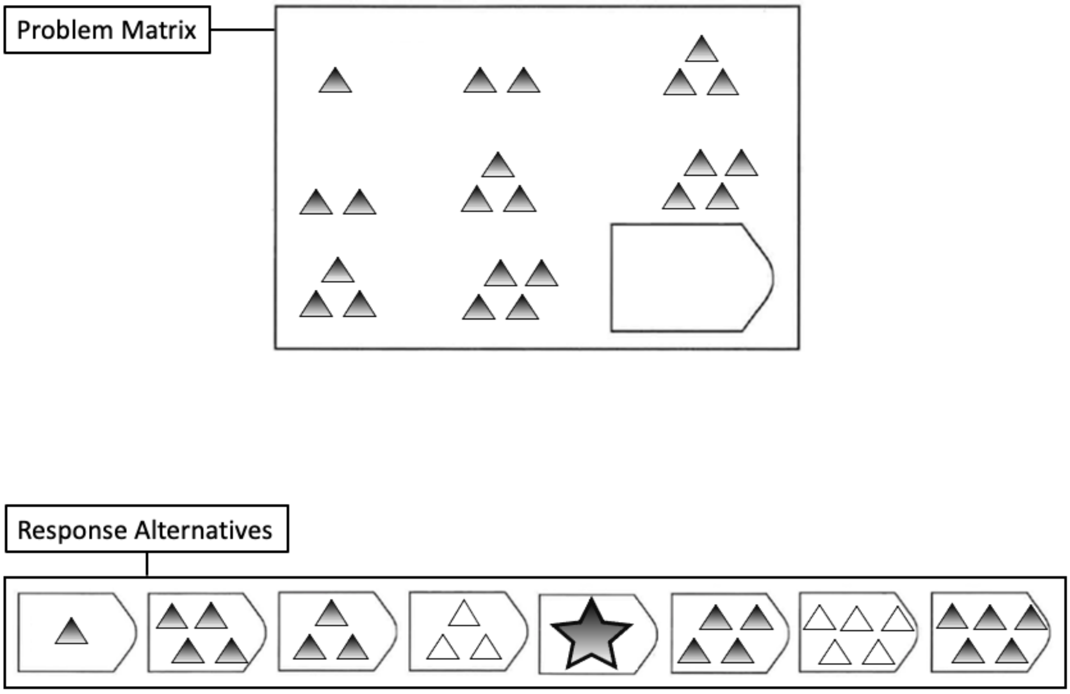
A special focus set on fluid intelligence, might provide further insight and help resolve at least parts of this giant conundrum. Fluid intelligence shows the strongest relation to a general intelligence factor (in models that account for it, Carroll, 2003 / Gustafsson, 1984) and can be described as a conglomerate of different reasoning abilities that appear in most intelligence models (e.g., Horn 1991; Spearman, 1904). Carroll (1993) who made substantial contributions in the field also stated that reasoning abilities are usually at the core of what is meant by the term intelligence.

Reasoning ability can be defined as the ability to solve unfamiliar or novel problems (Cattell, 1963; McGrew, 2006; Schneider & McGrew, 2012) but also can refer to a bundle of very similar abilities such as, detecting relations or patterns, identifying specific rules, drawing inferences, and solving abstract problems (Carroll, 1993). Reasoning ability is frequently used as an indicator of general intelligence (Roth & Herzberg, 2008) and a vast amount of research points toward the positive relation with scholastic achievement beyond for example motivational (e.g., Kriegbaum, et al., 2018) or personality factors (Laidra, et al., 2007). Consequently, this highlights the importance of truly understanding what is being studied with reasoning ability measures and what creates individual differences therein.

While the construct of human intelligence is still heavily discussed, how to measure reasoning ability is often not. Well-established reasoning ability measures are valid measures of fluid intelligence (Gustafsson, 1984; Kan et al., 2011; Schweizer et al., 2011) and the *Raven's Advanced Progressive Matrices* (APM; Raven, 2000) is seen as a state-of-the-art reasoning ability measure. Just as most reasoning ability measures, the APM consist of a large set of similar abstract problems / items the participant must solve. For each item the problem matrix (see Figure 1) shows eight geometrical figures with the ninth entry missing. The participant must identify the rules used to manipulate the figures from left to right or top to bottom in order to infer what the missing entry should look like. The task is then to select the correct figure to complete the set in the problem matrix from the presented response alternatives.

Figure 1

Example item for reasoning ability measures



Note. This item was designed to closely resemble the items of the Raven's Advances Progressive Matrices for illustrative purposes only and was not used in any experimental setting. Figure was adapted from Study 2 (von Gugelberg & Troche, *in preparation*).

Design choices and number of items might differ between frequently used reasoning ability measures, yet most rely on providing several response alternatives (for an exception see DESIGMA, Becker & Spinath, 2014), consequently making the measure adhere with the common multiple-choice format.

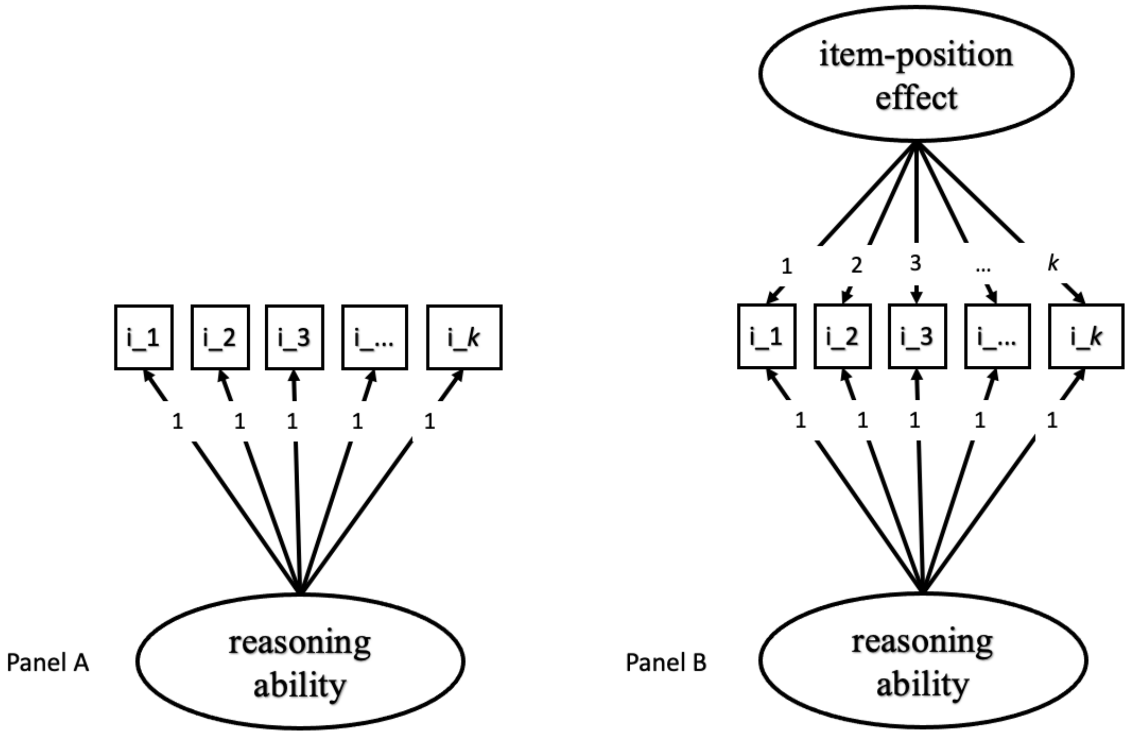
Another shared characteristic of reasoning ability measures is that only a small number of rules are applied to all items. These rules to manipulate the figures presented in the problem matrix are used separately, combined and /or applied to different elements in each figure of an item (Carpenter et al., 1990). Therefore, rules are repeatedly used over the course of a single testing.

Despite the highly similar item material, reasoning ability measures have frequently been found to be not as homogenous as assumed. A one-factor solution often failed to describe the data well (e.g., Dillon et al., 1981; van der Ven & Ellis, 2000; Vigneau & Bors, 2008). This lack of homogeneity implies that other factors are at play creating individual differences not necessarily related to the reasoning ability itself.

The item-position effect has been repeatedly named as the potential culprit for this lack of homogeneity and is the focus of this dissertation. The item-position effect points towards the phenomenon that a bifactor model (see Figure 2 for illustration), where a second latent variable capturing the increasing (true) item variance from the first to the last item of a reasoning test accompanying the latent variable representing reasoning ability, yields a better data description than a one-factor solution (e.g., Zeller, Krampen et al., 2017).

Figure 2

A one-factor model and a bifactor model for a reasoning ability measure with k items



Note. Panel A depicts a one-factor model, Panel B a bifactor model with two latent variables for a reasoning ability measure with a total of *k* items. Displayed fixations must be weighted by the standard deviation of the respective item. Fixations of the item-position effect are additionally divided by *k* for a linear increase or first squared and then divided by k^2 for a quadratic increase. Means and covariances of latent variables are set to zero for both models. Detailed explanations for model specification are provided in the Statistical Analysis subsection.

Improved model fit when the item-position effect is accounted for was found for several different reasoning ability measures, such as the APM (Ren, Schweizer et al., 2017), Cattell's (German adaption of Weiss, 2006) Culture Fair Test (Troche et al., 2016), Formann et al.'s (2011) Vienna Matrices Test (von Gugelberg et al., 2021) and Horn's (1983) sequential reasoning test (Ren, Gong, et al., 2017). The approach of implementing bifactor models by the means of confirmatory factor analysis was coined as fixed-links modeling (e.g., Schweizer et al., 2012) and since the (true) item variance increases from item to item, the effect was simply named the item-position effect while its underlying source was still unknown.

The item-position effect was shown to be of predictive value beyond reasoning ability itself as it served as a better predictor for verbal and math grades of secondary school students compared to the latent variable representing reasoning ability (Ren et al., 2015). These results suggest that the item-position effect adds to the predictive validity of reasoning tests for school grades above and beyond reasoning ability. Hence the item-position effect does not simply represent a method effect but individual differences that are psychologically meaningful and therefore must be better understood.

This dissertation aims to provide further insight on the item-position effect in reasoning ability measures, by taking a closer look at the learning hypothesis and the idiosyncratic test taking behavior of individuals.

The Learning Hypothesis

Currently the most plausible explanation for the item-position effect relies on different areas of research and points towards the possibility of individual differences in the ability to learn rules during the completion of a common reasoning ability measure. A very detailed analysis of rules was made by Carpenter et al., (1990) and that not only learning during test taking occurs (e.g., Carlstedt et al., 2000), but also individual differences therein occur was observed by Verguts and De Boeck (2000). The latter study also observed that the differences in learning rate seemed to be rule specific. Briefly summarized these studies indicate that an individual who can grasp the needed rules in the early stages of test completion, has an advantage when solving later items requiring the same rules compared to individuals who have difficulties learning certain rules. An individual who has difficulties learning certain rules will have to start the solving process over and over again with each item, possibly straining their resources and facing more opportunities for errors to sneak in.

Another group of researchers found the learning trajectories in the APM to be associated with item position but not item difficulty (Birney et al., 2017). This is especially interesting since most reasoning ability measures rely on progressive item difficulty, just as the APM, where item difficulty increases from item to item. Another empirical study (Zeller, Reiss et al., 2017) and simulation study (Schweizer & Troche, 2018) dissociated the item-position effect from item difficulty.

The assumption of learning, most likely as a consequence of repeatedly having to apply the same rules led Ren et al. (2014) to the learning hypothesis for the item-position effect. The learning hypothesis postulates, that individual differences in the ability to learn specific rules during test taking varies between participants, hence making the variance in response behavior grow larger throughout a test. And this increase in variance is what is captured by the added second latent variable with monotonically increasing factor loadings in the model. The then often found improved model fit (e.g., Schweizer et al., 2012) indicates a more comprehensive data description.

A high correlation between complex-rule learning tasks found by Ren et al. (2014) provided first evidence for the learning hypothesis. These results were later conceptually replicated by Schweizer et al. (2021). Yet simple correlations are not enough evidence to draw final conclusions. Especially since the item-position effect also showed high correlations with other factors. One study demonstrated a relation between the item-position effect and attentional control (Cowan, Fristoe et al., 2006), another with executive attention (Ren, Gong, et al, 2017), another with updating and inhibition (Ren, Schweizer et al., 2017) and even with several personality factors (e.g., Birney et al., 2017).

This highlights the necessity to further the understanding of what exactly elicits an item-position effect. In broader terms it also indicates that what truly happens during test taking has not yet been fully understood. There is a lack of research investigating individual differences during the test taking process and it is finally being noticed. For example, Birney and Beckmann (2022) point to the importance of accounting for differences in performance not only between participants but also possible changes within an individual during test completion (i.e., learning or experience).

The goal of this dissertation is just that; To provide more information on the item-position effect regarding the learning hypothesis, all while taking a closer look at the test taking process and individual differences therein.

Research Questions and Hypotheses

Relevant constructs for each study in relation to the item-position effect and the learning hypothesis are briefly explained below. For a more detailed explanation, including alternative hypotheses please refer to the cited articles (see Appendix A – C for full text).

In a first study (Study 1) we assessed whether the item-position effect and therefore indirectly the assumed rule learning is related to attentional control processes and individual differences therein (von Gugelberg et al., 2021). In a second study (Study 2), we used eye tracking measures to take a direct look at what happens during test taking. The goal was to see whether a specific strategy fostering rule learning could be identified and whether strategies in general are related to the item-position effect (von Gugelberg & Troche, *in preparation*). In a third study (Study 3) we put the learning hypothesis to the test with an experiment designed to disrupt the item-position effect during test taking (von Gugelberg et al., 2025).

Study 1: Attentional Control and the Learning Hypothesis

The dual mechanisms of cognitive control suggested by Braver (2012) entails two dissociable mechanisms of attentional control (Gonthier, Braver, et al., 2016). Proactive control refers to the mechanism of early selection and maintenance of goal-relevant information during task completion. This mechanism of control guides attention in anticipation of a challenging event, preparing the individual to successfully conquer it. Reactive control, as the name implies, indicates a late mobilization of attention initialized by a stimulus or event. Specific information is not anticipated to prepare processing in advance.

While Braver (2012) suggested that proactive control demands more resources, other studies found direct positive relations to fluid intelligence (e.g., Gray et al., 2003; Burgess & Braver 2010; Lu et al., 2016) and other indicators of mental ability (WMC, Redick, 2014; Richmond et al., 2015). For example, the study of Gray et al. (2003) manipulated the amount of interference in an n-back task. Results indicated that proactive control made individuals more resilient against interference since the goal of the task is actively maintained, indicated by stronger brain activity related to proactive control. Individuals with lower fluid intelligence were outperformed by individuals with higher fluid intelligence in the n-back task where the amount of interference was manipulated.

That higher fluid intelligence facilitates or enables the use of the more resource demanding proactive control was one conclusion drawn from the results of Gray et al. (2003) and Burgess and Braver (2010). We suggested an alternative explanation, where the opposite direction of effects take place.

Specifically, we proposed that an individual engaging in proactive control has an advantage when solving common reasoning ability measures where the same rules and similar stimuli are used repeatedly. And this advantage leads (at least to some degree) to higher reasoning ability scores. Proactive control allows an individual to select and maintain previously learnt information. In reasoning ability measures that maintained information would be the rules applied to solve previous items.

Taking the learning hypothesis into account, we assumed, that individuals engaging in proactive control would have the learnt rules readily available when solving the next item. This has their attention ideally biased to detect already learnt rules and possibly makes it less likely for them to miss an important detail or to lose trace of an already learnt rule. With each item solved, their advantage grows unlike individuals engaging in reactive control.

With reactive control every single item is first processed without prior gained knowledge about rules in mind. Only, and only when something akin to a previously used rule in the present item is detected, does this individual access the previously learnt information. For this process to be successful the rule must be identified accurately, and information retrieval must be correct. This creates several opportunities for things to go wrong, and such individuals seem to gain little that could work to their advantage during test completion.

To conclude, we predicted individuals engaging in proactive control to exhibit a larger item-position effect, due to their active maintenance of rules learnt compared to individuals engaging in proactive control. To test this hypothesis a task (AXCPT) that successfully distinguishes between the two mechanism of control was implemented (Gonthier, Macnamara, et al., 2016). With a latent profile analysis, groups with different reaction time patterns were identified and the relation of each group with the item-position effect was discerned.

Study 2: Strategy and the Learning Hypothesis

In 1978 Snow was able to identify individual differences in the solving process of reasoning ability measures and his research is finally generating some traction today. His discovery of two distinguishable strategies was conceptually replicated in for example, eye tracking data (Vigneau, et al., 2006), verbal protocols (Jarosz, et al., 2019) or short questionnaires (Gonthier and Thomassin, 2015).

The two strategies identified are response elimination and constructive matching. With the latter, individuals spend a lot of time inspecting the problem matrix, most likely identifying the different rules applied to each entry to then mentally construct the missing entry. After mentally constructing the solution, the individual then selects the matching entry from the presented response alternatives. Hence its name, constructive matching. With response elimination the individual eliminates non-viable response alternatives step by step, until a solution is found.

To investigate the mentioned strategies, collecting eye tracking data is an objective method that offers an abundance of information about the solving process of each individual for every item. Several studies have taken advantage of these benefits (e.g., Vigneau et al., 2006; Hayes, et al., 2015; Laurence et al., 2018) and the Toggle Rate has been used as the most straightforward and intuitive indicator for strategy use in reasoning ability measures.

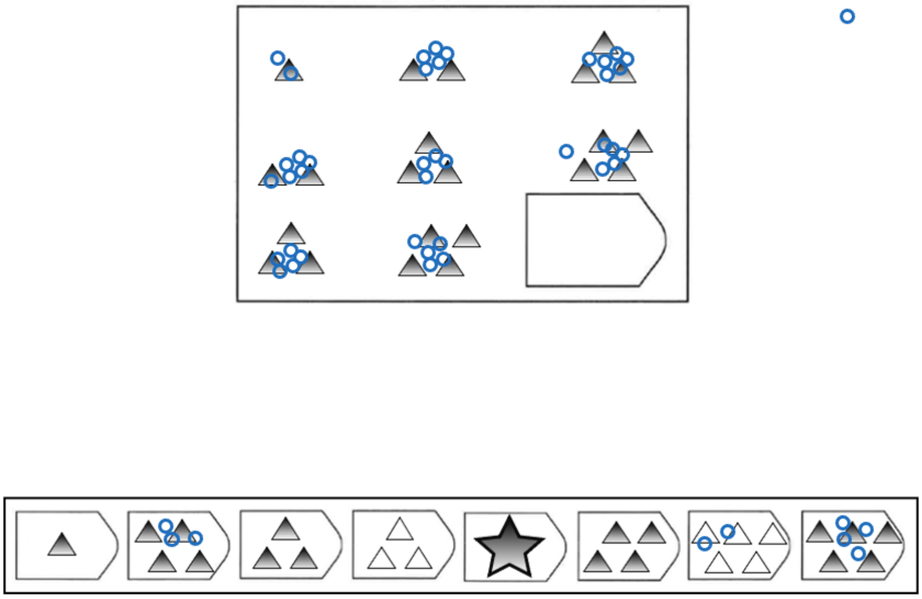
The Toggle Rate directly translates the observations made by Snow (1978) that switching frequently from problem matrix to response alternatives to eliminate non-viable responses is indicative of response elimination and fewer switches with more time spent on the problem matrix is indicative of constructive matching. In terms of eye tracking measures, it is the sequence of fixations between interest areas that is being quantified.

A fixation refers to those times when the eyes stop looking around and stand still to take in detailed information. The looking around or scanning occurs between fixations and is referred to as a saccade (SR Research, 2016). A Toggle occurs when a fixation on the problem matrix is followed by a fixation on the response alternatives or vice versa. The chances of Toggles occurring, increases with item latency. When an individual spends more time solving an item, this individual has more time to toggle between problem matrix and response alternatives. Therefore, the Toggle Rate takes item latency into account. The Toggle Rate quantifies the number of times an individual toggled from looking at the problem matrix to the response alternatives (and vice versa) divided by total time spent on the item.

Hence, a low Toggle Rate is indicative of constructive matching, since the individual spends most of the time inspecting the problem matrix and only exhibits few toggles to the response alternatives and back (see Figure 3). A high Toggle Rate occurs when an individual shows many toggles between response alternatives and the problem matrix, which is in line with the strategy of response elimination (see Figure 4). By using the Toggle Rate as an indicator of strategy, strategy is operationalized as a continuum, where constructive matching makes up one end, and response elimination the other.

Figure 3

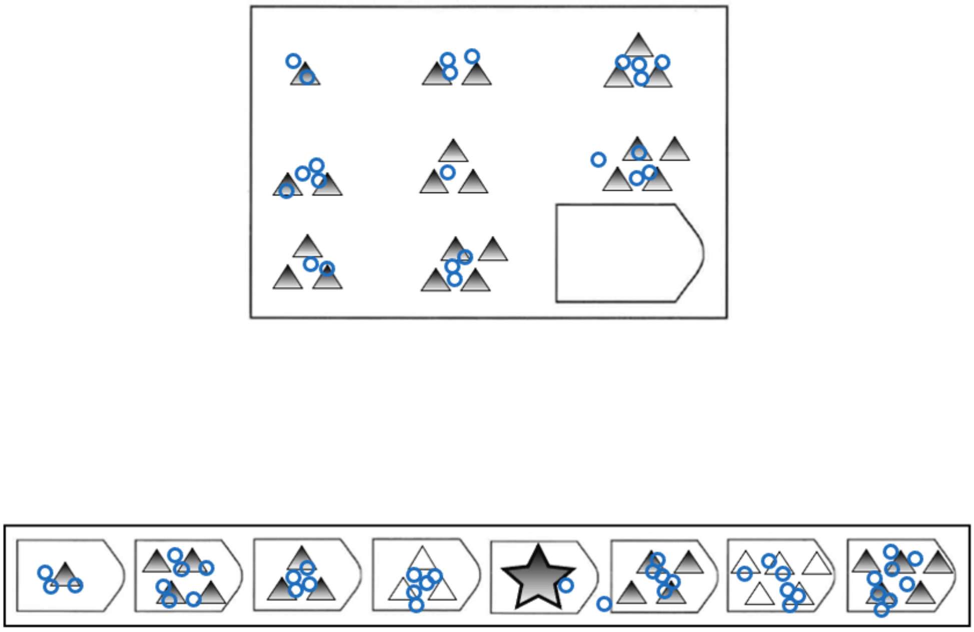
Simplified example of fixation pattern of the constructive matching strategy.



Note. A fixation is indicated by a small blue circle. Many fixations on the problem matrix and few fixations on the response alternatives is a typical fixation pattern for the constrictive matching strategy. Sequence of fixations cannot be derived from this figure but constitutes necessary information to determine number of toggles.

Figure 4

Simplified example of fixation pattern of the response elimination strategy.



Note. A fixation is indicated by a small blue circle. Fewer fixations on the problem matrix and many fixations on all the response alternatives is a typical fixation pattern for the response elimination strategy. Sequence of fixations cannot be derived from this figure but constitutes necessary information to determine number of toggles.

The reasons why different individuals rely on different strategies is an ongoing debate. Some name test properties (e.g., Raden & Jarosz, 2022) and others individual differences in mental resources (e.g., Gonthier & Thomassin, 2015; Jarosz et al., 2019; Li et al., 2022). Bethell-Fox et al. (1984) and Snow (1978) concluded from their work that strategy use is influenced by the interrelationship of mental resources and the item properties perceived by the individual. They observed that whenever the capacity to hold rules and manipulate objects was exceeded, the individual would switch from constructive matching to response elimination. A similar shift in strategy was observed by Gonthier and Roulin (2020). Participants progressively shifted from constructive matching towards response elimination. Work by Vigneau et al. (2006) found no shift in strategy, but rather an initial difference between subjects on what strategy they engage in. Their data and other results (e.g., Jastrzebski et al., 2018) report a positive correlation between constructive matching and reasoning ability, leaving the conundrum unresolved.

The question whether a shift in strategy occurs in a similar fashion for all participants (Gonthier & Roulin, 2020), whether it is the initial ability of the individual that decides what strategy an individual applies (Vigneau et al., 2006) or whether it is actually both, remains unanswered. This led to the first objective of this study. By using the same fixed-links modeling approach coined by Schweizer (2006) for the item-position effect a bifactor model can be fit to the Toggle Rate. Such a model allows to account for an innate difference in strategy (as observed in Vigneau et al., 2006) and concurrently a potential shift in strategy (as observed in Gonthier & Roulin, 2020).

The second and relevant objective for this dissertation is concerned with the applied strategy and rule learning. From the early descriptions of Snow (1978) and Bethell-Fox et al., (1984) I derived constructive matching to be the more expedient strategy to learn the rules necessary to solve an item correctly. With this strategy individuals not only spend a lot of time inspecting the problem matrix but also seem to engage in a methodical analysis of the stimuli (Snow, 1978). Therefore, I assumed constructive matching to foster rule learning, and any rule learnt should facilitate solving subsequent items, since rules are repeatedly used in common reasoning ability measures. With response elimination on the other hand, it is less likely for individuals to learn all the underlying rules to an item correctly. The individuals spend less time on the problem matrix and are mainly concerned with eliminating non-viable response alternatives.

This led to the assumption that under the premise of the item-position effect reflecting individual differences in the ability to learn rules during test taking, the relation between the item-position effect and constructive matching should be a positive one. Individuals applying constructive matching during test taking systematically analyze the rules of the problem matrix and then mentally construct the correct solution. Hence, they are more actively engaged with the underlying rules of an item compared to individuals relying on response elimination who dismiss response alternatives step by step.

To analyze such idiosyncratic behavior during test completion, we tracked participants eye movements during the completion of the APM. For the first objective we ran a confirmatory factor analysis with fixed-links to investigate the Toggle Rate. The goal was to see whether the previously identified individual differences can be depicted by a bifactor model. This approach allows for potential individual differences in the overall strategy applied (as in Vigneau et al., 2006) but also for a shift in strategy (as in Gonthier & Roulin, 2020).

The same statistical approach was used for the APM score data to discern whether an item-position effect occurred. The main objective of the study could only be investigated, if an item-position effect was present in the data. Under the premise that rule learning underlies the item-positing effect, the latent variable depicting the item-position effect should be related to the latent variables depicting innate and / or change in Toggle Rate. Since lower Toggle Rate indicates more constructive matching, and it is constructive matching that I assumed to foster rule learning, I predicted a negative correlation between the latent variables depicting the item-position effect and Toggle Rate.

Study 3: Rule disruption and the Learning Hypothesis

As the learning hypothesis is to date a plausible theory on what underlies the item-position effect, the goal of Study 3 was to test this hypothesis as directly as possible with an experimental design. We assumed that, if it is ad-hoc rule learning underlying the item-position effect, a sudden change of the underlying rules within a sequence of items, should disrupt the item-position effect.

Hence, the goal was to create a test that specifically favored the emergence of an item-position effect, in order to then disrupt the item-position effect with an experimental manipulation. Assuming rule learning as the underlying source of the item-position effect, repeated use of only few rules for a larger set of items should create an advantage for the assumed underling ability to learn rules during test taking and an item-position effect should emerge. When such a sequence of items is followed by a sequence of items with different rules, the item-position effect should be disrupted, since the prior learnt rules no longer provide any insight about the newly introduced rules.

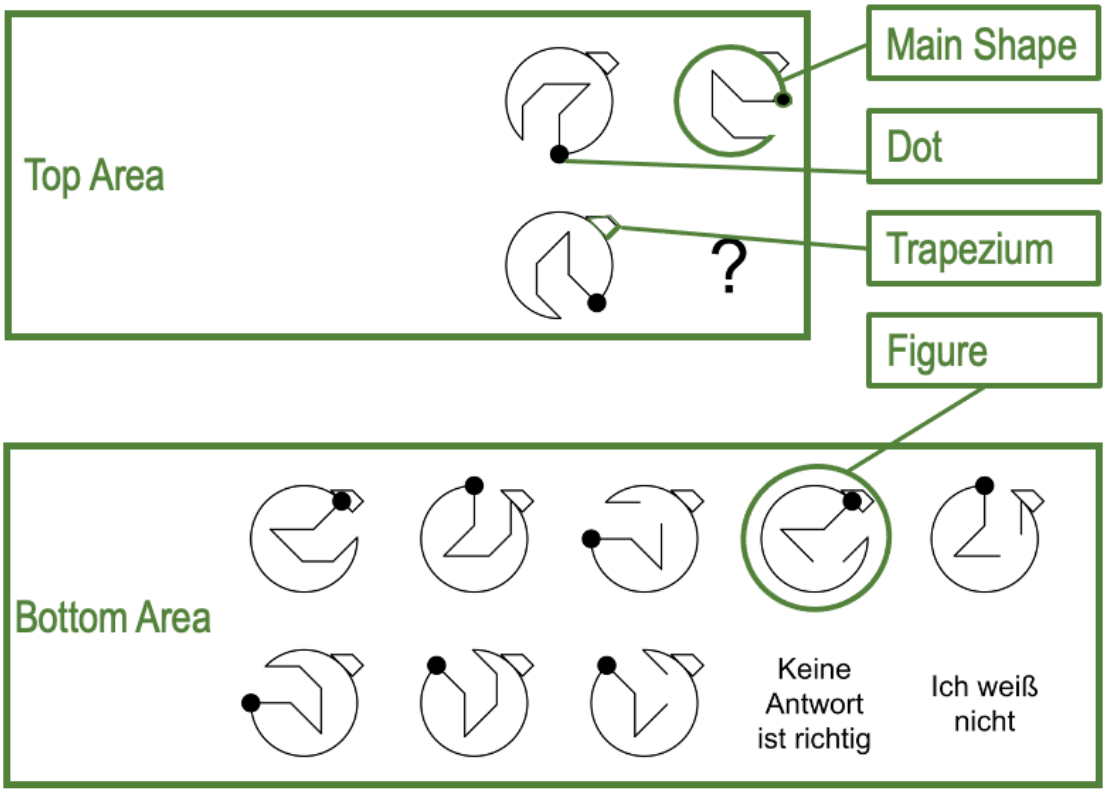
With this premise, a reasoning ability test was needed that allowed for hypothesis driven item creation, yet in a familiar enough design, that results can be transferrable to common tests such as the APM. I assumed this to be possible with items generated in the

IMak R package (Blum & Holling, 2018). The figural analogies that can be created with the *IMak* R package seemed to fulfill all requirements. Items can be created according to several different rules, distractors and overall item generation and difficulty parameters showed solid results (Blum et al., 2016). And as with most common reasoning tests, the implemented rules had to be identified it the top area (see Figure 5) relying on the information given in the first column or the first row to then extrapolate the missing entry signified by a question mark.

For the study I created a set of items based on only one rule. This rule entailed the movement of an element in the figure (see Figure 5). By applying this rule to different elements of the figure simultaneously a variety of items with varying difficulty could be created. An additional set of items was created using two different rules, that entailed the subtraction of straight lines and mirroring of the figure. After launching a pilot study where participants completed all items, two Tests of Figural Analogies (TFA) were created. The pilot study provided insight on item difficulty and completion times. This allowed for an informed item selection for the two final versions of the TFA. A detailed description of the pilot data and item selection for the final test creation is given in the Method section under the subsection Implemented Reasoning Ability Measures.

Figure 5

Item created with the IMak package to illustrate items of the Test of Figural Analogies (TFA)



Note. Figure adapted from Study 3 (von Gugelberg et al., *2025*). Things added to the item for illustrative purposes are colored green.

Briefly reiterated, the TFA in the continuous rule condition consisted of 24 items, where only the movement rule was used. Item difficulty increased throughout the test as the rule was first applied to only one element, followed by items where the rule was applied to two elements and thereafter to all three elements. This continuous use of the same rule we assumed would be highly beneficial for rule learning and hence should elicit an item-position effect. The second TFA in the discontinuous rule condition was identical to the first, yet upon the 18th item the two different rules where a straight line is subtracted, and the figure itself is mirrored were applied to the last six items. This sudden change of rules should disrupt an item-position effect.

With the creation of these two TFA the overreaching hypothesis was, that in the continuous rule condition a typical item-position effect emerges, while in the discontinuous rule condition the effect is disrupted preventing configural invariance between conditions.

More specifically we assumed that in the continuous rule condition two latent variables would be needed to describe the data best. One for reasoning ability and an additional one for the item-position effect. Hence a bifactor solution should outperform a one-factor solution if indeed an item-position effect emerged. This should also be the case for the discontinuous rule condition, yet we assumed a three-factor solution to outperform both one-factor and bifactor solutions with two latent variables, since through the disruption of rules the initial item-position effect should disperse and a new one (with the presentation of new rule) should emerge. Hence the need for a third latent variable to describe the data adequately.

Methods

This section provides an overview of participants, materials and the statistical analyses for all three studies. Section headers include parentheses to indicate if a section is only relevant for certain studies. Additionally, I took the liberty of omitting details that are not central to the overall results and conclusions drawn here. Further details can be found in the cited articles (also see Appendix A – C). Where applicable I mentioned what Section or Tables in the cited articles the additional information can be found in.

Participants

University students participated for credit while other participants could enter in a raffle to win a prize. Study 1 (N = 210; von Gugelberg et al., 2021) and Study 2 (N = 210; von Gugelberg & Troche, *in preparation*) were conducted in a laboratory whereas the experiment for Study 3 took place online (N = 403; von Gugelberg et al., 2025). All three studies were approved by an ethics committee. All participants gave written and informed consent.

Implemented Reasoning Ability Measures

A brief summary is given for the established reasoning ability measures only, and any deviations from the manual are mentioned specifically. For the newly created reasoning ability measure (developed for Study 3), item generation, rule selection, pilot data and final item selection is outlined in detail.

The Vienna Matrices Test (Study 1)

The Vienna Matrices Test (VMT; Formann et al., 2011) consists of 18 items. Item construction is similar to Figure 1 with a question mark in the bottom right entry. Participants were instructed to choose one out of eight possible response alternatives. As directed by the manual, no time limit was set. Age-stratified IQ scores were calculated, by adding up all correct answers for each participant and transforming them according to the representative sample given in the manual.

Raven's Advanced Progressive Matrices (Study 2)

The Raven's Advanced Progressive Matrices (APM) is intended for high aptitude population (Raven & Raven 2003). An item for illustration purposes was created (Figure 1). The sole adaption of item material for the study was to present all eight response alternatives in a single line to facilitate tracking eye movement (as displayed in Figure 1).

Participants received the instruction given by the manual and as prescribed completed two example items followed by the 36 items. A time limit of 30 minutes was set. For the analysis every single score (1 = correct answer / 0 = false answer) was used.

Tests of Figural Analogies (Study 3)

The *IMak* package in R (Blum & Holling, 2018) was used to create two specific Tests of Figural Analogies (TFA). Albeit the small differences between figural analogies and typical matrices, studies showed that items created with the IMak package show satisfactory convergent validity with other common measures for reasoning ability (Blum & Holling, 2018). Since new items were created specifically to test our hypothesis, item creation is explained in detail. The thereafter described pilot data provided information for the final item selection and test creation.

Item Creation.

While the package offers a set of different rules, only three rules were implemented. The rules used in the study could be applied to different elements within a figure, i.e. main shape, dot and trapezium (see Figure 5). The IMak package allows the figure itself to be mirrored (Rule 1), straight lines of the main shape to be removed (Rule 2) or single elements to be moved (Rule 3). Rule 3 can be implemented clockwise and counterclockwise to the trapezium or the opening of the figure. Additionally, the degree of movement can be specified, for example a 45° or 90° movement. The movement of the dot along the edges of the opening can be defined and again, direction of said movement and the number of edges the dot passes can be specified. Examples for the movement of each element can be found in Figure 6.

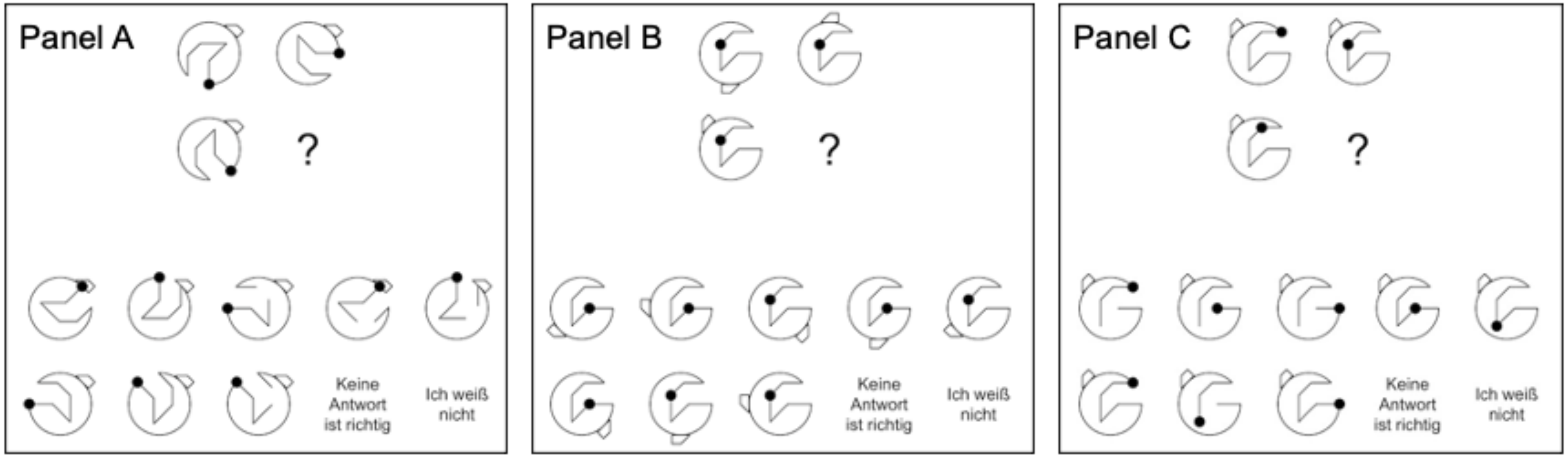
The goal was, to apply the movement rule (Rule 3) with increasing difficulty to the different elements. Starting with simply moving only one element in an item (as in Figure 6), created easy items, and the rule applied is very understandable. An increase in difficulty was achieved by implementing two (or three for the most difficult items – example given in Figure 7, Panel A) movement rules to an item. For example, the trapezium was moved clockwise by 45° and the main shape was moved counterclockwise by 90°. The possibility to apply the same rule to multiple elements in an item, I assumed would provide the best scenario for rule learning to occur and therefore would be most likely to provoke an item-position effect.

If this repeated use of the same rule is followed by items with very different rules, a disruption of the item-position effect, if it's truly based on rule learning should occur. Therefore, items with the other two selected rules were created. That is, a line was removed (Rule 1) and the main shape (Rule 2) was mirrored. An example for this item is given in Figure 7, Panel B.

To check whether the sudden change of rules truly could disrupt the item-position effect a control group was necessary, where the item-position effect was not disturbed by new rules. Hence the goal was to create one set of items, where the item-position effect develops from beginning to the end of a test, and one set of items, where the item-position effect can develop, but is then disrupted with the introduction of new rules. Ideally both item sets would be identical in the beginning and only deviate when the new rules are introduced in the experimental condition. With this in mind, several items were created, and a first pilot study with the newly created items was run.

Figure 6

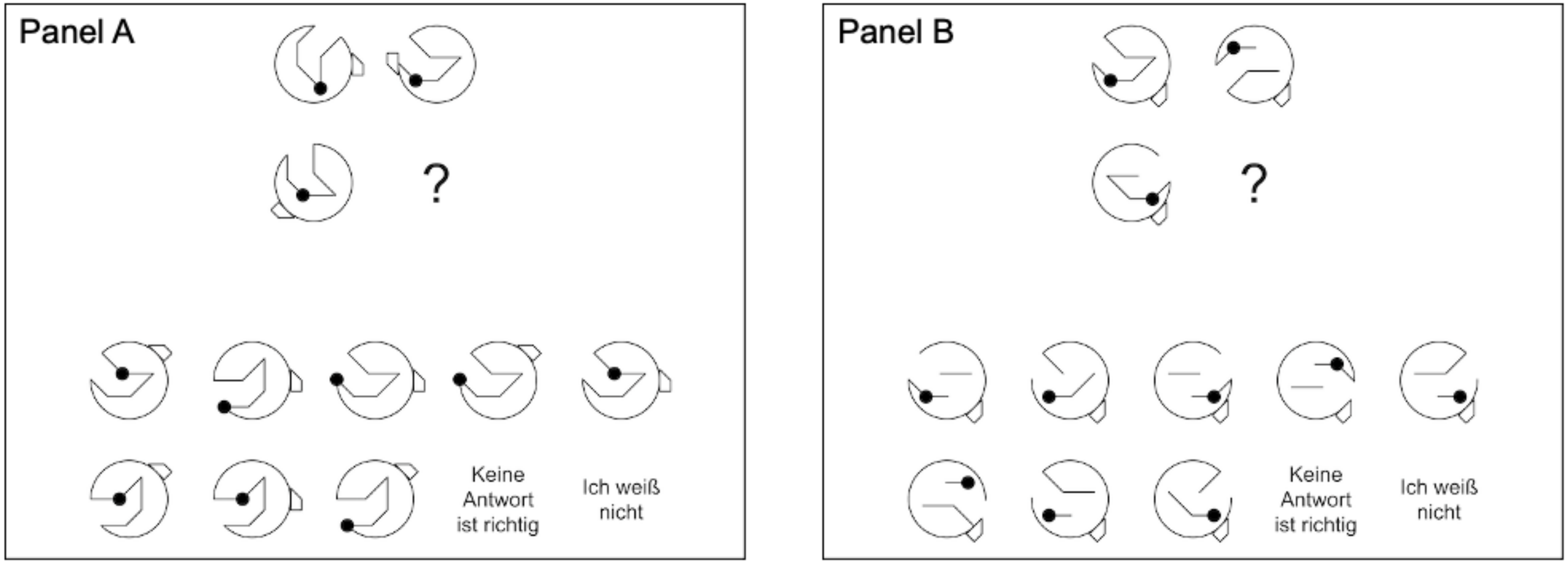
Example items for each movement rule created with the IMak package


Panel A
?
Keine Antwort ist richtig
Ich weiß nicht
Panel B
?
Keine Antwort ist richtig
Ich weiß nicht
Panel C
?
Keine Antwort ist richtig
Ich weiß nicht

Note. Panel A shows the movement of the main shape, Panel B the movement of the trapezium, Panel C the movement of the dot.

Figure 7

Examples for rules used in the last six items of the continuous and discontinuous rule condition



Note. Panel A shows an item were main shape, trapezium and the dot are moved. A typical example for the continuous rule condition. Panel B shows an item from the discontinuous rule condition where a line was removed, and the main shape was mirrored.

Pilot Study.

The pilot study allowed to examine whether item difficulty increased with the simple addition of a rule as predicted by the creators of the IMak package (Blum et al., 2016) and also allowed to explore whether unintended ceiling or bottom effects emerged. The pilot study included instructions and practice items as recommended by Blum and Holling (2018). Practice items were only based on the movement rule, and the rule was only applied to one element. Practice items were then followed by 6 one-rule items (as in Figure 6), 9 two-rule items (movement rule was applied to two elements), 12 three-rule items (as in Figure 7, Panel A), and 9 line removal – mirrored items (Figure 7, Panel B). This resulted in a total of 36 items (see IMak-Full in Figure 8) for the pilot study. Item sequence can also be inferred from Figure 8.

In the pilot study 32 participants completed the 36 items of the IMak-Full. Thereof, 14 specified themselves as male, 17 as female. Mean age was 31.31 (SD = 10.68) and on average 25.19 (SD = 6.93) items were answered correctly, while the lowest score was 11 and one participant answered all 36 items correctly. Item difficulty and standard deviation for all items are displayed in Figure 8. As intended an increase in item difficulty can be seen in Figure 8, with the orange polynomial line depicting a downward trend.

Nonetheless some items showed a very large difference from the preceding and following items regarding their item difficulty. These items are highlighted in Figure 9. A closer inspection of the highlighted items revealed, that in some cases the response alternatives had especially good lures, which seemed to have increased item difficulty unintentionally (e.g., see item 15 in Figure 10). Other items had overall very similar response alternatives making the items easier as intended (e.g., see item 13 in Figure 11). Since response alternatives are generated by an algorithm of the *IMak* package, manual inspection and pilot data is very important and, in this case, revealed some suboptimal items. The highlighted items were substituted with new items created with the exact same rules but better response alternatives.

Experience showed that usually 18 items are enough to elicit an item-position effect. This meant, that we wanted the first 18 items to be identical for both conditions and also elicit an item-position effect. For the experimental condition the item-position effect should be disrupted by the introduction of new rules, hence six items with the two other rules (i.e., line removal, mirroring) where selected to follow the first 18 items. With the change in rules, this was labelled the discontinuous rule condition.

As a control, a continuous rule condition was created using the exact same first 18 items but continuing with the same rule for the additional six items. Hence all 24 items of the continuous rule condition were created with only the movement rule, providing ideal conditions for the item-position effect to occur.

Implemented Tests of Figural Analogies.

Final item selection from the pilot study is depicted in Figure 12. The first 18 items were used in both conditions with the noted substitutions for items 7, 13, 15 and 17. From the 3-rule items an additional six items were selected for the continuous rule condition. For the discontinuous rule condition six out of the nine items with the line removal and mirroring rule were selected. This resulted in two sets of items, each set containing 24 items, whereof the first 18 are identical.

For the study participants were instructed to choose the corresponding figure among the presented response alternatives in the bottom area or select the most accurate verbal response (“No response is correct” and “I don’t know”). All instructions were adapted from the material provided in Blum and Holling (2018). Verbal feedback from our pilot study indicated that the instructions were understood clearly. All participants first received a general instruction about an item and its elements (similar to Figure 5), accompanied by a generic instruction on how to solve an item. This was followed by the same three practice items used in the pilot study, were only Rule 3 (movement of an element) had to be applied. These items had to be solved correctly to proceed to the next item. The practice items were followed by 18 items all relying on Rule 3. Participants randomly assigned to the continuous rule condition were then presented an additional six items also created with Rule 3. Participants randomly assigned to the discontinuous rule condition were presented six items created with Rule 1 and 2 (i.e., main shape mirrored, and straight line subtracted). Neither condition had a time limit nor received feedback on any of their responses after completing the practice items.

Figure 8

Item characteristics, their difficulty (Pi) and standard deviation in the pilot study

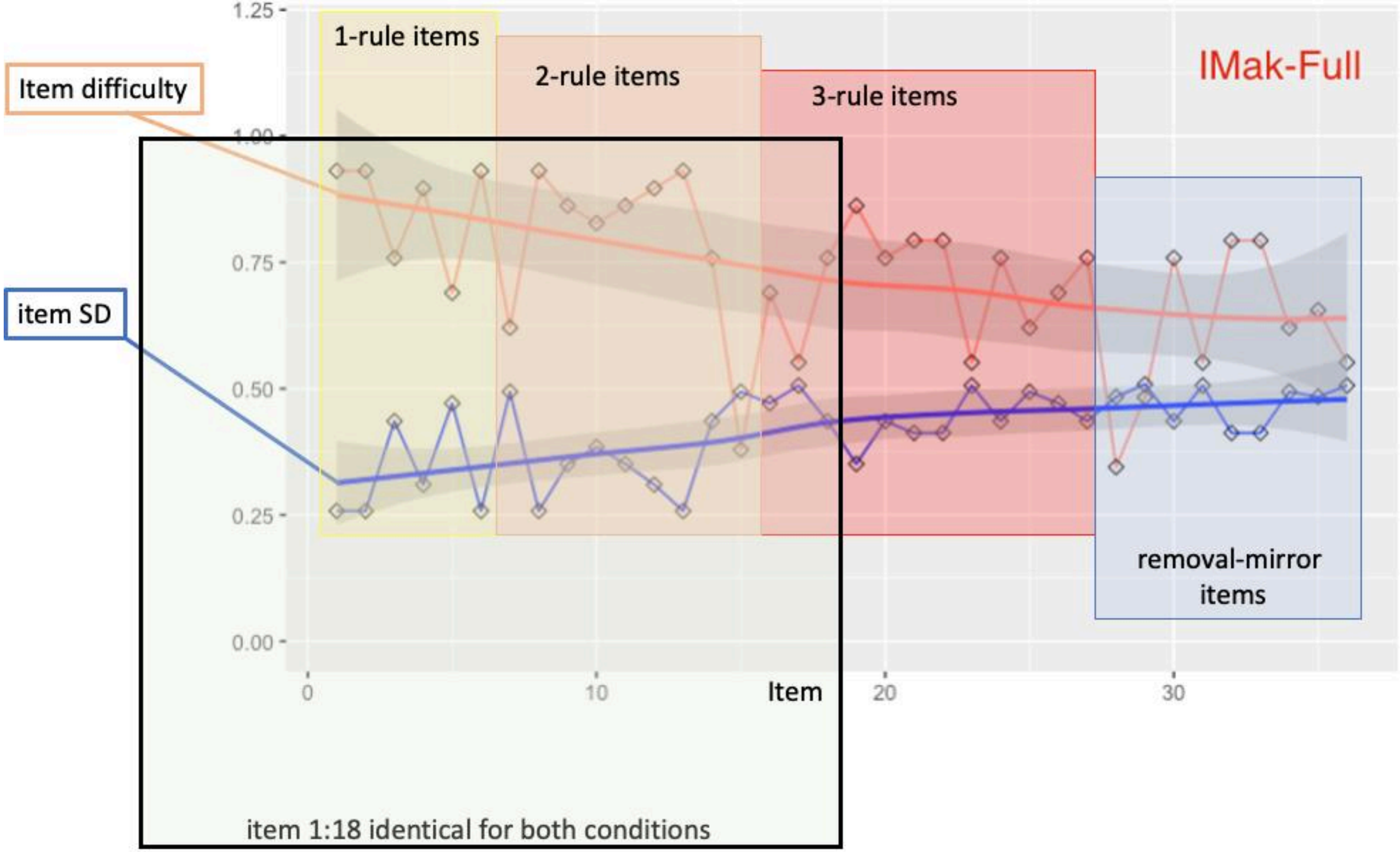
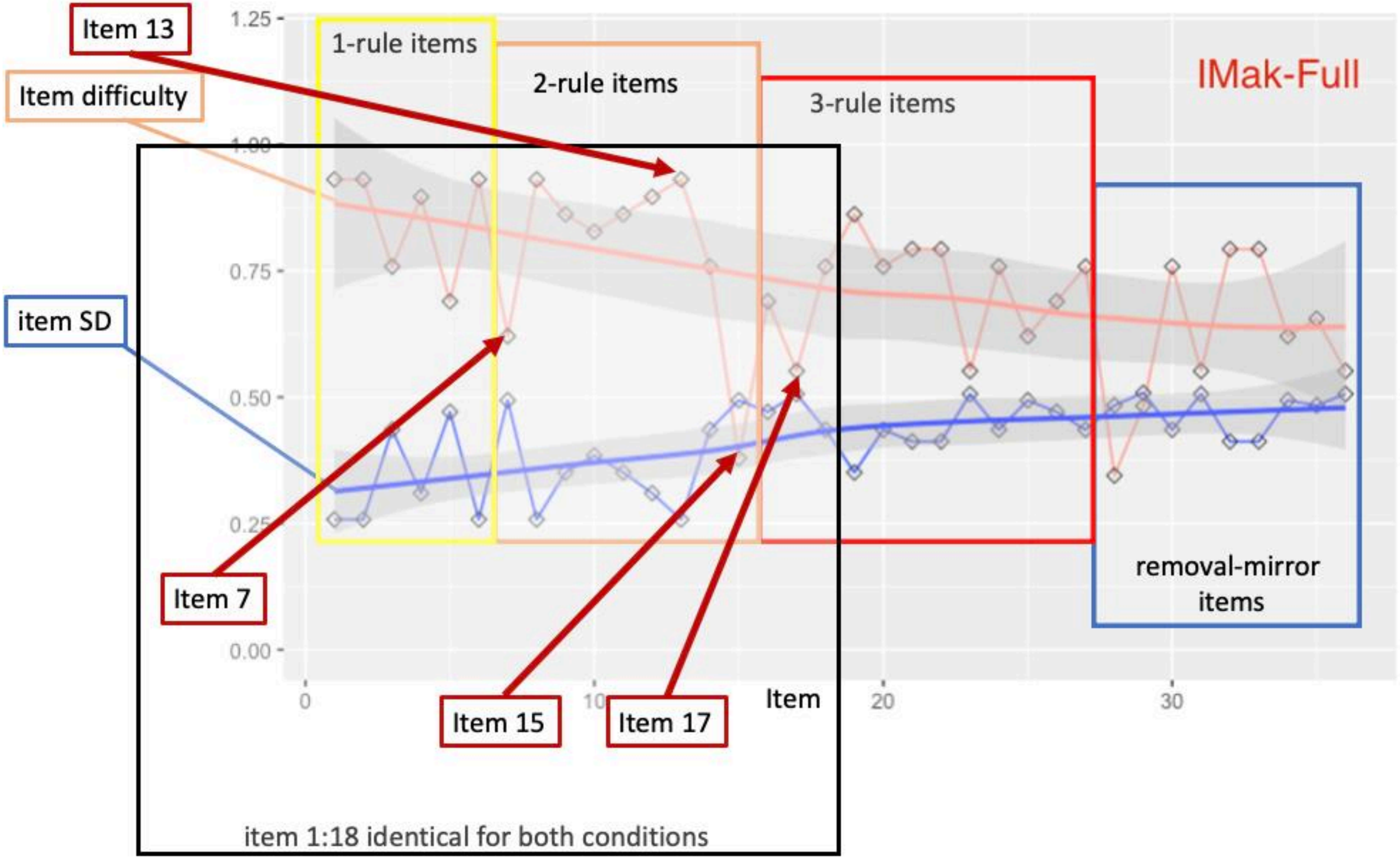


Figure 9

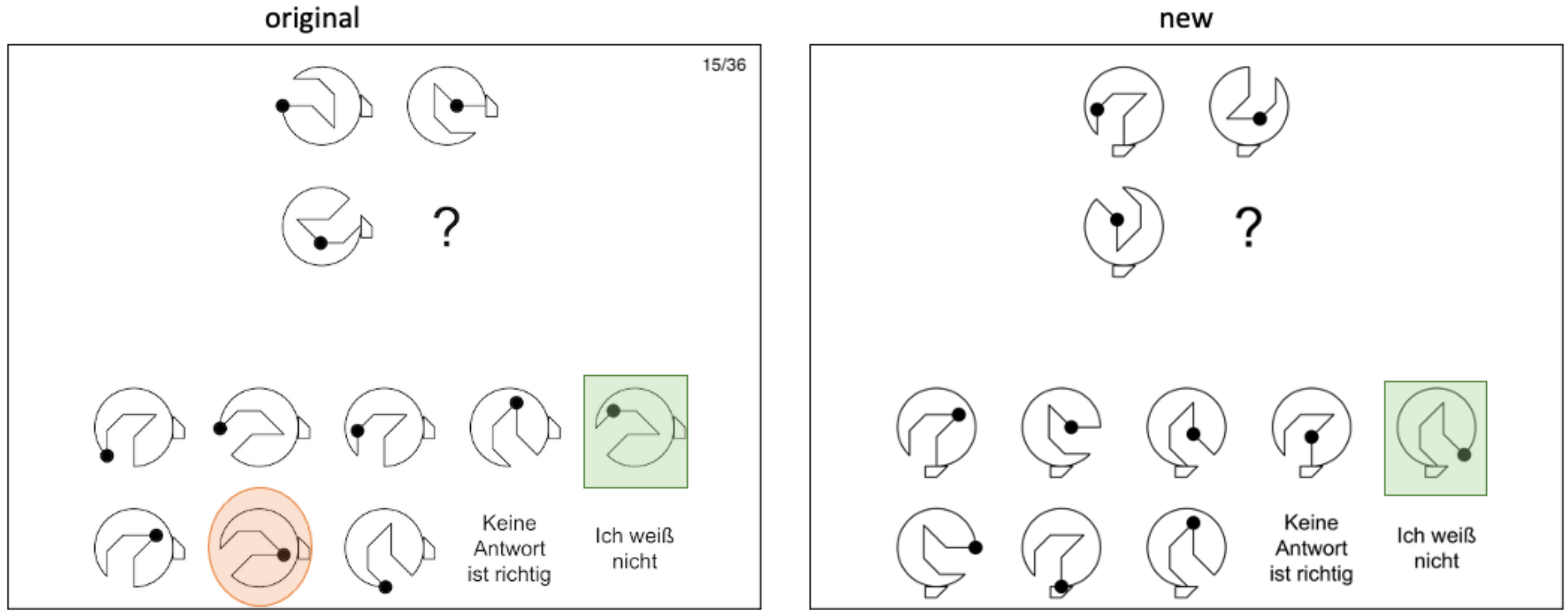
Item characteristics, their difficulty (Pi) and standard deviation in the pilot study, with suboptimal items pointed out



Note. Marked items showed problematic distractors and deviated in item difficulty from their neighboring items.

Figure 10

Example of the original item with suboptimal lures and the new item used as substitute



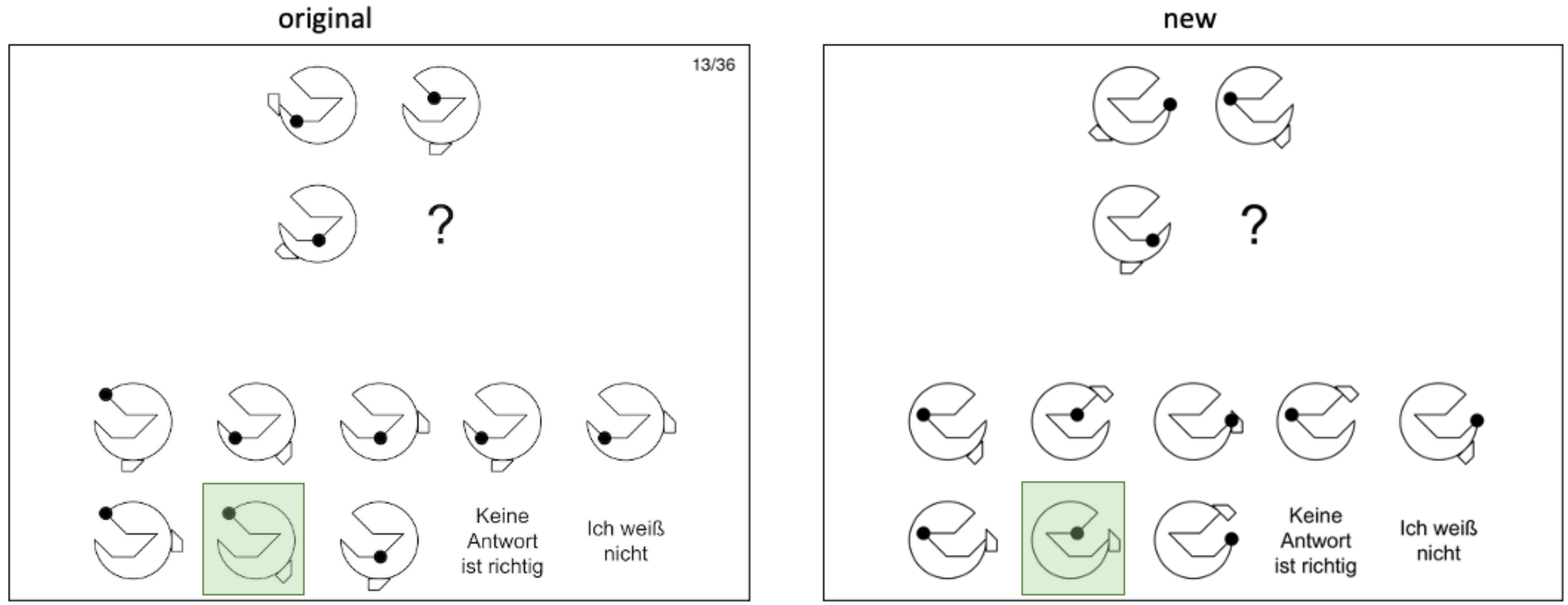
original

new

Note. Correct answer is marked with a green square in both items. In the original item, 16 participants selected the circled answer in red, which in this case was not correct. It seems this was a very good lure, making the item more difficult.

Figure 11

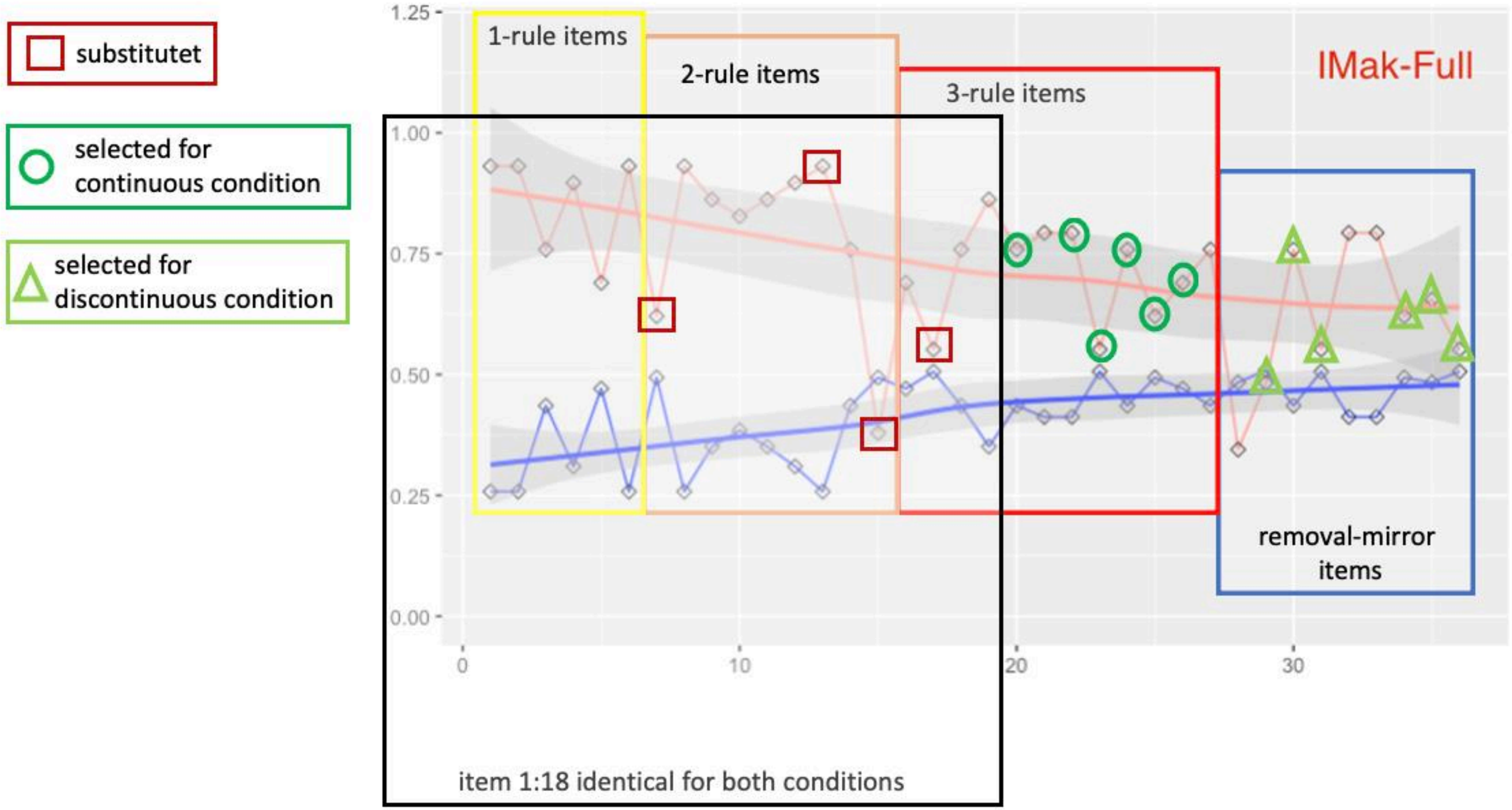
Example of the original item with suboptimal lures and the new item used as substitute



Note. Correct answer is marked with a green square in both items. In the original item, the lures in the first row are very similar, making the item easier.

Figure 12

Display of pilot items, their selection and substitution for the final two versions of item sets



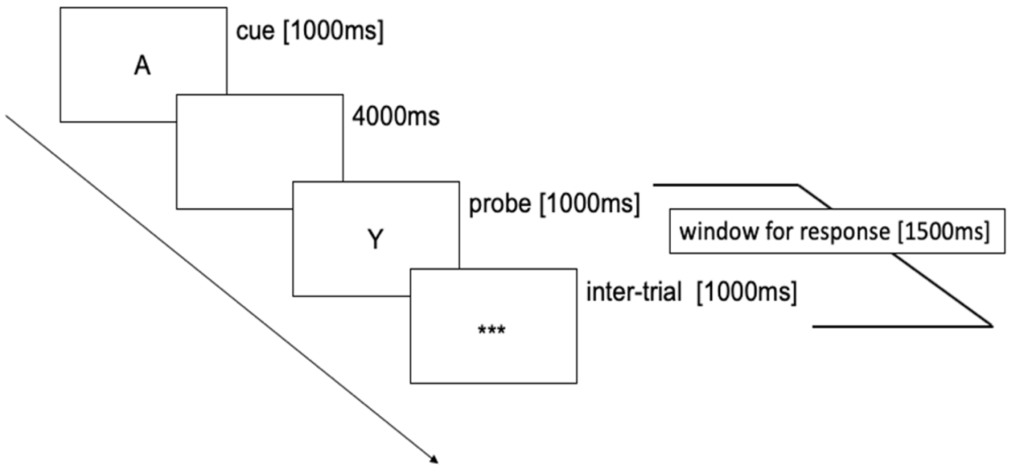
Continuous performance task (Study 1)

In the AX-Continuous Performance Test (AX-CPT), participants see a cue letter followed by a probe letter (see Figure 13). Four different combinations of cue and probe letter exist. That is, in the AX-condition, the cue letter “A” is followed by the probe letter “X”. This is the target condition, and participants are tasked with pressing a designated key whenever this target condition appears. All other letter combinations are non-targets and require participants indicating them as such by pushing another designated key. The non-target conditions are abbreviated as BX-condition, BY-condition and AY-condition. The letter “B” always indicates any letter except “A” as cue, “Y” any letter except “X” as probe, and the letters “A” and “X” represent themselves as cue or probe respectively.

Reaction time (RT) differences between these four conditions have been interpreted as indicators for reactive or proactive control (e.g. Paxton et al., 2008; Braver et al., 2001; Redick 2014; Gonthier, Macnamara, et al., 2016).

Figure 13

Simplified display of one trial of the AY condition of the AX-CPT.



Note. Duplicate of Figure 1 in von Gugelberg et al., (2021)

An individual fully engaging in proactive control keeps the relevant cues in mind, and as soon as they appear prepares for the correct response. This means in both conditions where the cue is a non-target letter (BX- and BY-condition) individuals prepare for the non-target response immediately, resulting in very short RT’s. Additionally, since the probe letter is not relevant in these conditions when relying on proactive control, no difference in RT between conditions should emerge. With the other two conditions (AX-, AY-condition) the target cue can be followed by a target or non-target probe and the AX-condition is presented four times as often as any other condition. Hence, RTs should be notably slower compared to the BX- and BY-condition.

The dual mechanism of control theory indicates that individuals relying on reactive control wait for the probe to appear, before forming their response. Looking only at the probe would indicate, that the two conditions with non-target probes (AY-, BY-condition) should result in similar RTs, since no retrieval of the cue is necessary to form a response. When the probe is a target (AX-, BX-condition), RTs should be somewhat slower and more ambiguous, since different information (was cue “A” or NOT “A”) must be retrieved.

This indicates that, several differences in RT mark the use of proactive or reactive control. The mentioned differences (i.e., differences in RT pattern, NOT in subtracted values) were then used to discern whether the overall RT pattern reflected proactive or reactive control.

Toggle Rate (Study 2)

Details on calibration / validation process, setup in the laboratory, procedure etc. can be found in the method section of von Gugelberg and Troche (*in preparation*). Only a brief explanation for the calculation of the Toggle Rate is given here.

Monocular eye data of every individual for each item was checked for drift. Drift is a systematic shift of all fixations in a certain direction. This can occur when an individual moves slightly, despite the used chin-forehead rest. Such movement can result in all fixations landing on blank areas on the screen. If drift was detected during the visual inspection of the fixations, all fixations for an item were moved in cohesion. After this raw data inspection, all fixation sequences for every participant for each item were extracted to calculate the Toggle Rate.

A Toggle is defined as a fixation on the problem matrix followed by a fixation on the response alternatives or vice versa. Fixations outside of the mentioned areas were recoded with the area of the previous fixation. Meaning, if a participant first fixates on the problem matrix, then stares at empty white space on the screen (possibly thinking, see Figure 3, top right corner for such an example of a fixation) and then shows the next fixation on the response alternatives, it will be counted as a Toggle. Without the recoding, this type of scan path would not be counted as a toggle, although it qualifies as an alternation between problem matrix and response alternatives. After adding up all Toggles per item, the Toggles were divided by the corresponding item latency. This procedure results in a Toggle Rate for each item of every individual. The Toggle Rate was used for further analysis.

Statistical Analysis

This section provides a brief overview of the different methods used. Detailed information can be found in the cited articles. All analyses were run in R (R Core Team, 2020).

Fixed-links Modelling (all studies)

Fixed-links modeling was coined by Schweizer (2006) and as the name implies, relies on fixed links (factor loadings) when investigating the factorial structure through the means of a confirmatory factor analysis (CFA) of a test with homogenous item material. The fixation of factor loadings is theory driven and allows one to extract more than one latent variable from the same set of observed variables. For the item-position effect, whether it is the individual difference in the ability to learn rules during test completion or not, the assumption holds, that the contribution to the observed variance of the latent variable depicting the item-position effect increases from the first to the last item. Therefore, the factor loadings are fixed to increase from the first to the last item. Currently it is unclear whether a linear or quadratic increase for the latent variable depicting the item-position effect is preferable, and often model selection at this point is no longer theory driven but data driven, i.e., the model describing the data best is selected.

The bifactor structure, where one latent variable depicts reasoning ability and another the item-position effect (illustrated in Figure 2 Panel B), must not only outperform a one-factor solution, but also yield significant variance parameters for both latent variables. If the model fit of the one-factor solution outperforms the bifactor model solution, data description is inferior in the latter and there is no reason to opt for a less parsimonious model.

Simplified, the three models described in Table 1 are fit to the data, if a bifactor model outperforms the one-factor solution, an item-position effect was successfully detected in the data. This was the case for several different reasoning ability measures (e.g., Ren, Gong, et al., 2017), and such a consideration of the item-position effect also could be shown to improve validity of a reasoning ability measure (Schweizer et al., 2012).

Table 1

Adaption of Table 1 in von Gugelberg et al., (2025)

Overview of calculated models and details regarding the specifications.

Model	Structure	Latent Variable(s)	Fixation of Factor Loadings
Model 1	one-factor model	reasoning ability	1 or no fixation
Model 2	bifactor model	reasoning ability	1 or no fixation
		item-position effect (linear course)	$\frac{i}{k}$
Model 3	bifactor model	reasoning ability	1 or no fixation
		item-position effect (quadratic course)	$\frac{i^2}{k^2}$

Note. All factor loadings were additionally weighted by SD$_i$ as link function. Letter *i* refers to the respective items position, *k* refers to the total number of items (i.e., 24), SD to the standard deviation. In Study 1 and 2 no fixation was set for the reasoning ability latent variable, in Study 3 they were fixed to 1.

While the gist of the approach is simple, a lot of details during model estimation and evaluation of fit are important to consider and implement. Regarding model estimation, it first must be noted, that the data analyzed is binary. This requires either specific estimators and thresholds estimated from tetrachoric correlations, or as suggested by Schweizer (2013) a threshold free approach with maximum likelihood estimation can be implemented. The latter is preferable, since tetrachoric correlations must not only provide an estimate of the relation between two variables, but also bridge the difference between a binomial and normal distribution (Schweizer et al., 2015). This leads to disproportionally large distortions on the estimated thresholds especially in tail areas (values close to one or zero) which in turn disproportionally influence the estimated tetrachoric correlations (Schweizer et al., 2015).

These difficulties can be circumvented with a probability-based covariance matrix (calculation does not use an asymptotic function) and an additional link function to account for the switch from binary variables to the normal distribution (Schweizer, 2013). The probability-based covariance matrix used for threshold free approach can be directly calculated with a function in the *bindabox* package (von Gugelberg, 2022) or indirectly implemented with *lavaan* (Rosseel, 2012), by adding specific commands. Since no threshold is used for the estimation a link to the binary data must be directly included in model specifications. Schweizer (2013) recommended using the items standard deviation as said link. Therefore, every item in the model specification is fixed with its link (i.e., standard deviation), making the name of the approach (fixed-links modelling) of a descriptive nature.

For each item, in addition to its standard deviation as a link, the theory driven fixation is needed. Reasoning ability can be assumed to be equal across all items (all factor loadings fixed to 1 – Study 3) or to vary across items (factor loadings freely estimated – Studies 1 & 2). For the item-position effect either a linear or quadratic increase is modelled. Here fixations should always lead up to the final value of 1, as recommended by Schweizer (2009). This can be achieved by implementing the formula in Table 1 or calculated directly with functions of the *bindabox* package (von Gugelberg, 2022).

Also, from a statistical point of view, it is feasible to assume that the latent variable depicting reasoning ability to be independent from the one capturing the item-position effect, since it allows to set the correlation between the two to zero. This prevents the variances from overlapping, which can lead to difficulties in model estimations. Contrary to latent growth models no mean structure is assumed and therefore, no intercepts (means) of the latent variables are estimated.

To evaluate model fit, all common fit indices are considered (for details see von Gugelberg et al., 2025). Additionally, the estimated variance parameters are inspected. If the estimated variance parameters of a bifactor model do not significantly contribute to the model, there is no need for such a latent variable in the model, since it does not explain anything. The variance is tested for significance one-tailed, since negative values are in this case theoretically impossible and would indicate grave model misspecification.

The size of the estimated variances parameter (and their standard errors) is influenced by the size of the chosen factor loadings. Therefore, comparing these parameters in a model can only provide meaningful information after they were scaled according to the eigenvalue scaling method (Schweizer & Troche, 2019).

Latent profile Analysis (Study 1)

Testing the main hypothesis, that the item-position effect would be more pronounced in individuals engaging predominately in proactive control, came with several obstacles. Since it is the difference in RT and not the RT itself that provides information about the type of cognitive control a simple correlational analysis of RT did not seem feasible. A direct subtraction of RTs to obtain the mentioned RT differences is problematic, since it would be unclear what kind of individual differences the subtraction points to (for more detailed information on this issue see Draheim et al., 2019). Also, RTs are consistently related to higher reasoning ability scores (e.g. Der & Deary, 2017), rendering direct correlations between RTs of conditions and the item-position effect mute.

We assumed that different differences in RT should lead to clearly distinguishable *RT patterns*. Hence, we needed a method that can identify different patterns in RTs without the need for calculating RT differences. To identify possible groups of individuals engaging in proactive or reactive cognitive control a Latent Profile Analysis (LPA) seemed to offer what we needed. The LPA allowed us to detect different groups without relying on RT differences or coercing certain structures (e.g., assuming a certain number of groups) based on prior research.

In R the package *tidyLPA* (Rosenberg et al., 2019) applies an expectation–maximization algorithm to identify latent profiles. In a first step RTs were aggregated for each condition of every participant resulting in four mean RTs per participant. Then four different types of models with varying number of groups were fit to the data. Each model could have either equal or varying variances and zero or varying covariances. With each model type a group structure of two up to eight groups were tested. This resulted in 32 different solutions for the aggregated RT data. Taking several fit indices into account the best solution was then identified by an analytic hierarchy process (Akogul & Erisoglu, 2017).

The resulting groups from the best solution where then used for any further analyses. More detailed information about model types and model selection can be found in the Methods and Results section and Table 2 in von Gugelberg et al. (2021).

Multilevel Model (Study 1)

To statistically test RT differences between multiple groups without relying on RT differences or fully aggregated data a Multilevel Model (MLM) seemed the most feasible approach. After identifying the ideal number of groups to describe the RT patterns, the grouping variable and RTs per condition were analyzed. Different RT patterns between groups would be reflected in cross-level interactions, hence we calculated a Slope-as-Outcome model[1]. Within this model a cross-level interaction would be reflected in different slopes between groups, meaning different RT differences between groups and conditions.

Models were calculated with *lmerTest* (Kuznetsova et al., 2017) and Restricted Maximum Likelihood estimation (REML) as recommended by McNeish (2017). To compare all conditions between (3x3 intercepts, 3x6 slopes) and within groups (3x6) a total of nine

[1] Complete equation of the slope-as-outcome model: $RT_{ij} = \gamma_{00} + \gamma_{01}GroupB + \gamma_{02}GroupC + \gamma_{10}AY + \gamma_{20}BX + \gamma_{30}BY + \gamma_{11}AY{:}GroupB + \gamma_{21}BX{:}GroupB + \gamma_{31}BY{:}GroupB + \gamma_{12}AY{:}GroupC + \gamma_{22}BX{:}GroupC + \gamma_{32}BY{:}GroupC + \varepsilon_{ij} + \upsilon_{0j} + \upsilon_{1j}$. With *i* indicating the individual within a Group and *j* the Group, υ_{0j} the random effects of the intercept, υ_{1j} the random effects of the slope, ε_{ij} the residual variance.

Slope-as-Outcome models were calculated, where group and condition was re-leveled to provide the necessary information. More detailed information on the analysis can be found in the Methods and Results section and Tables 3, 4, and 5 in von Gugelberg et al. (2021)

Results

Key findings for the overreaching research question will be briefly reiterated from each study including the relevant details for the general understanding of results. Additional details can be found in the referenced articles.

Study 1: Attentional control and the Learning Hypothesis

The LPA identified three groups. Descriptive details on the groups can be found in Table 2. Group C, also the smallest group identified had significantly lower IQ scores compared to the largest Group A. Group B is somewhat in the middle and IQ scores do not differ from the other two groups.

Table 2

Duplicate of Table 1 in von Gugelberg et al., (2021)

Mean IQ, standard deviation (in parentheses) and reaction times (RT in milliseconds) in the four AX-CPT conditions for the full sample and subsamples identified by the latent profile analysis.

	VMT raw scores	IQ scores [1]	RT_{AX}	RT_{AY}	RT_{BX}	RT_{BY}
Full Sample (N = 210)	13.69 (3.06)	98.40 (14.34)	408 (99)	507 (100)	393 (139)	384 (129)
Group A (n = 114)	14.19 (2.92)	100.75 (13.62)	357 (30)	445 (40)	305 (34)	307 (31)
Group B (n = 67)	13.61 (2.93)	97.98 (13.70)	416 (48)	532 (55)	418 (56)	400 (53)
Group C (n = 29)	11.90 (3.30)	90.07 (15.77)	594 (130)	693 (88)	683 (110)	654 (115)

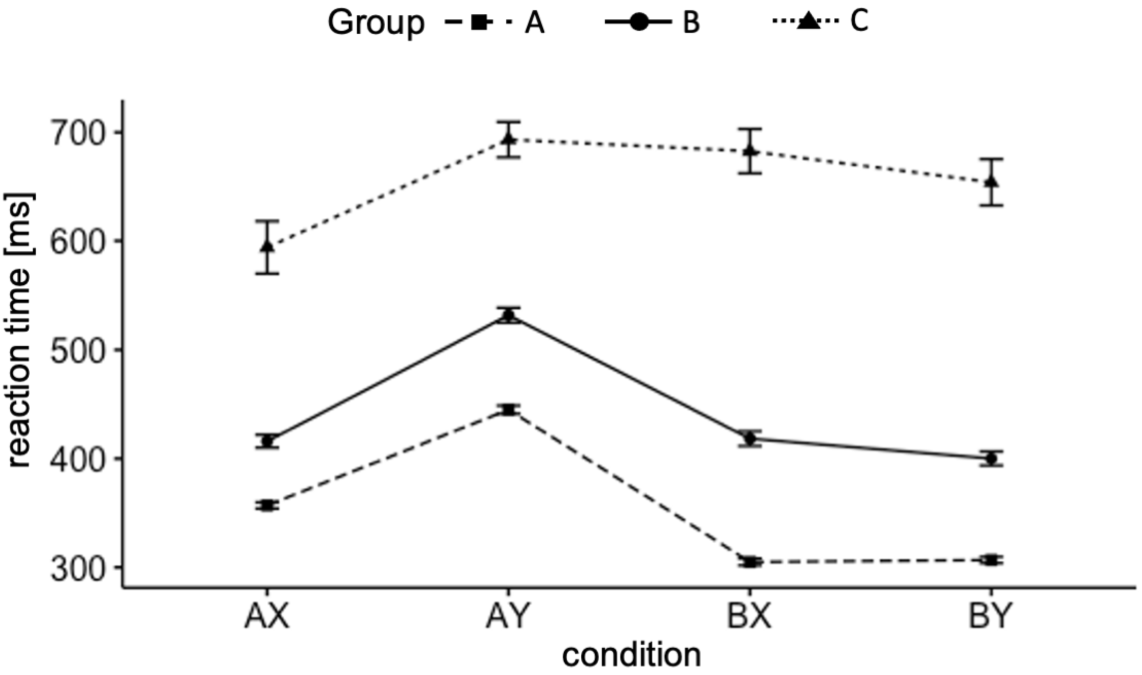
[1] Information about age for one participant is missing in the full sample and Group B, since IQ calculations were based on age-based norms, and no age was provided by one participant.

Specific details about group characteristics, noteworthy elements in their RT patterns and the analysed Slope-as-Outcome model are described in the section *7.2. Group characteristics* and Tables 3-5 in von Gugelberg et al. (2021).

Briefly summarized, Group A showed the most stereotypical pattern of proactive control. The RT pattern of Group C was most consistent with reactive control. Group B seemed to be somewhere in between, exhibiting behaviour consistent with either mechanism of control. Additionally, all RTs of Group C were significantly slower compared to all other RTs, and RTs of Group B were significantly slower than those of Group A. The observed RT pattern is illustrated for each group in Figure 14.

Figure 14

Observed reaction time pattern and standard errors in the four conditions of the AX-CPT for the three groups identified by latent profile analysis.



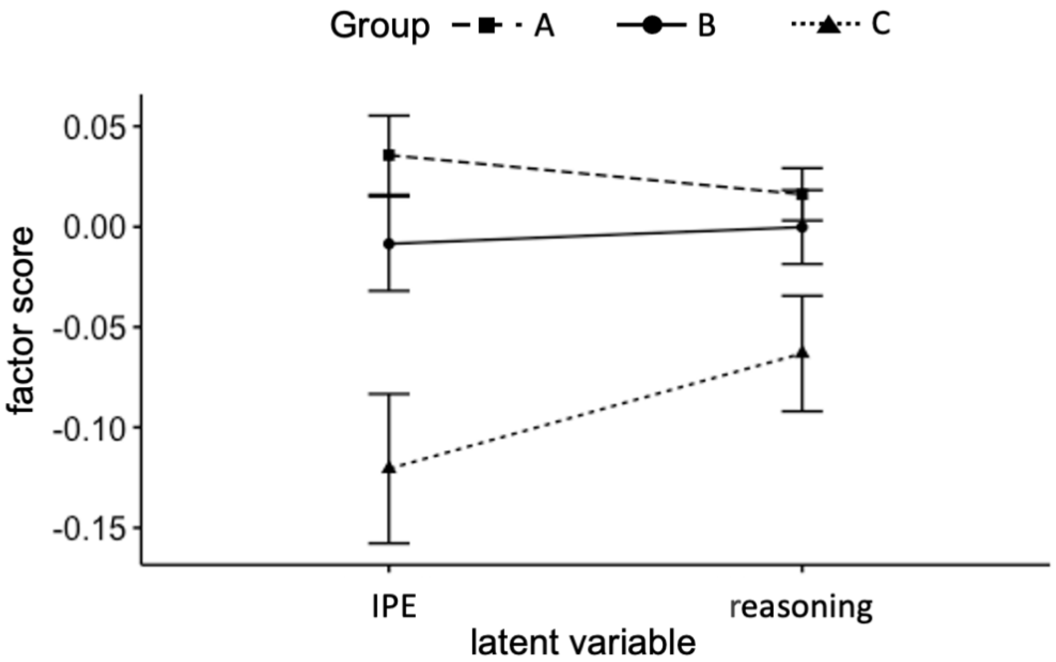
Note. Duplicate of Figure 2 in von Gugelberg et al., (2021).

After successfully identifying an item-position effect in the VMT (for details on model fit etc. see Table 6 in von Gugelberg et al., 2021), factor scores of said model were extracted for each participant. With two-tailed *t*-tests, factor scores for the latent variables reflecting the item-position effect and reasoning ability between the three groups were compared (Table 7 in von Gugelberg et al., 2021). No difference in reasoning scores between the three groups emerged. The item-position effect was only significantly different between Group A and C (see Figure 15).

This indicates that the item-position effect was more pronounced in Group A compared to Group C and that Group A engaged strongly in proactive control and Group C rather in reactive control. Also, when the item-position effect was accounted for in a model, no difference in reasoning ability emerged between groups.

Figure 15

Factor scores on the latent variables representing the item-position-effect (IPE) and reasoning ability in the Vienna Matrices Test for the three groups identified. Error bars represent standard errors.



Note. Duplicate of Figure 3 in von Gugelberg et al., (2021).

Study 2: Strategy and the Learning Hypothesis

Since not all participants completed all items (see Table 3) within the given time limit of 30 minutes only the first 27 items were used for the analysis. This excluded 3 participants from the full sample. Detailed reasoning behind this decision can be found in the results section of von Gugelberg and Troche (*in preparation*).

Table 3

Number of participants completing items at the indicated positions in the APM

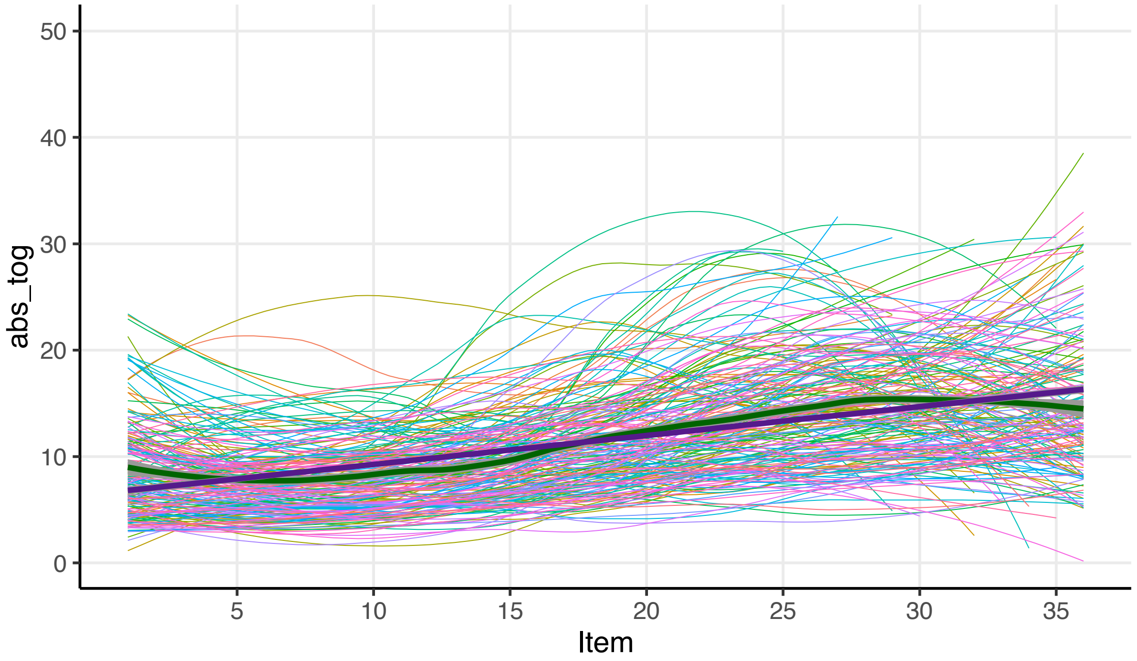
Item Position	26	27	28	29	30	31	32	33	34	35	36
N	210	207	206	204	199	194	190	186	177	169	164

Summarized, I decided this cutoff would allow to detect any shift in strategy and provide enough items to identify an item-position effect. Additionally, the bias of speediness, or potential time management skills would be minimal since participants within this sample took on average of 15 minutes to solve the 27 items. All participants (n =12) that took 25 or more minutes to solve the 27 items answered 16 to 25 items correctly, placing them nicely within the average of the sample. Participants (n = 9) with completion times just below 7 minutes to complete the 27 items, scored between 3 and 21. While the range of scores is quite large, these participants completed all 36 items. Indicating that they were not influenced by the time limit, but rather worked through the APM quickly.

For the analyzed sample the absolute numbers of Toggles increased throughout test completion (Figure 16). Item latencies also increased (Figure 17). With the increase in item latencies being steeper than the increase in absolute number of Toggles, the resulting Toggle Rate decreases from item to item (Figure 18).

Figure 16

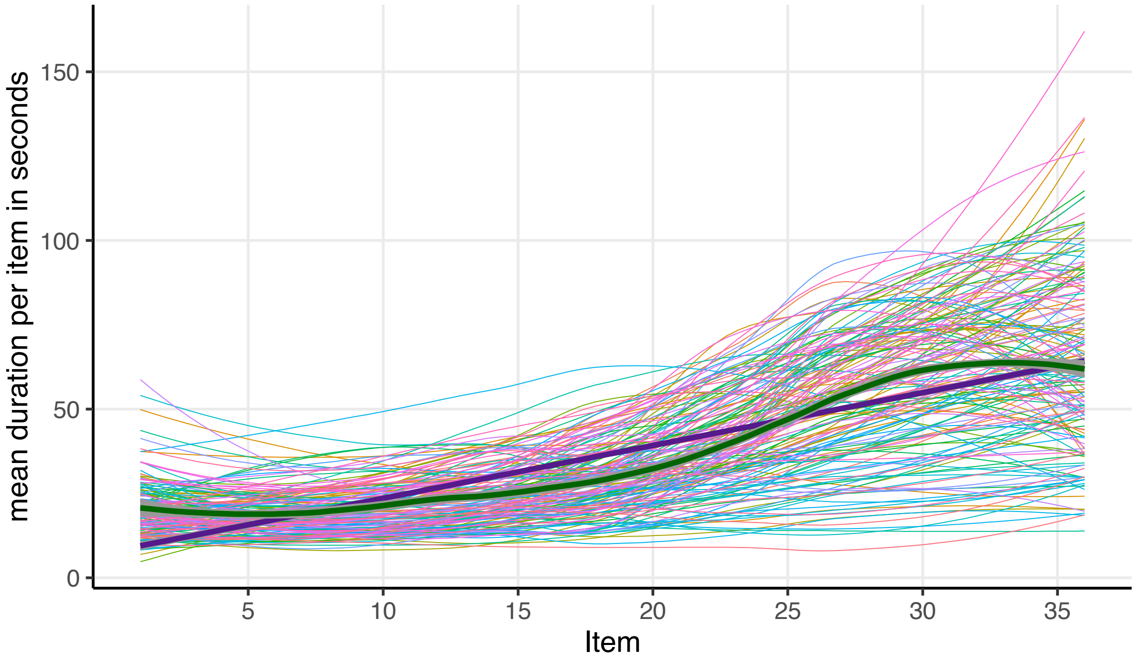
Number of Toggles for all 36 items



Note. Local polynomial (green) and linear (purple) regression line for the analyzed sample and colored lines for each participant.

Figure 17

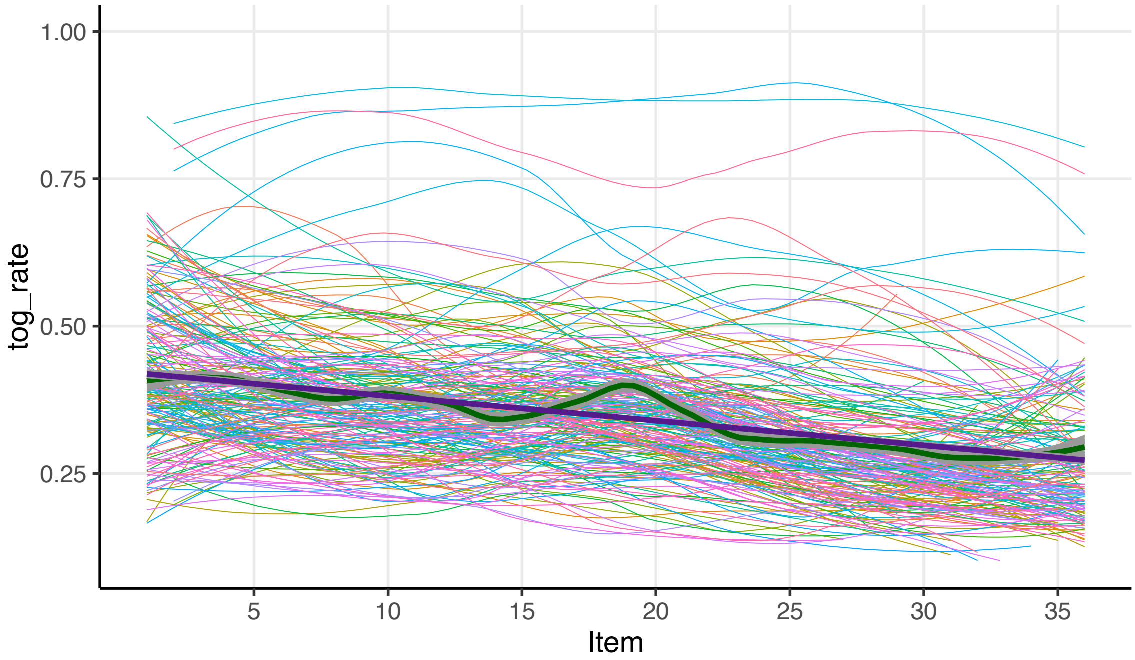
Item latencies for all 36 items



Note. Local polynomial (green) and linear (purple) regression line for the analyzed sample and colored lines for each participant.

Figure 18

Toggle Rate for all 36 items



Note. Local polynomial (green) and linear (purple) regression line for the analyzed sample and colored lines for each participant.

For both the Toggle Rate and the APM scores a bifactor solution with a linear increase provided the best data description for the first 27 items (see Table 1 in von Gugelberg & Troche, *in preparation*). To test the hypotheses a final model with the two bifactor models of Toggle Rate and APM scores was calculated. With the correlations in Figure 19 being somewhat difficult to interpret, an additional Figure was created to aid interpretation (Figure 20).

Figure 19

Correlations of the two bifactor models of Toggle Rate and APM scores.

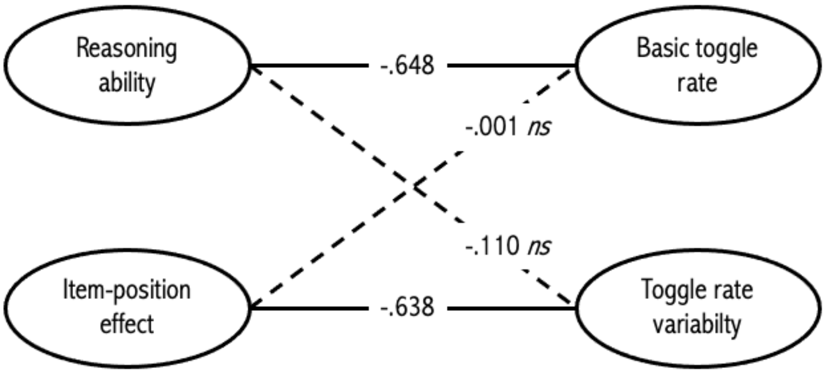
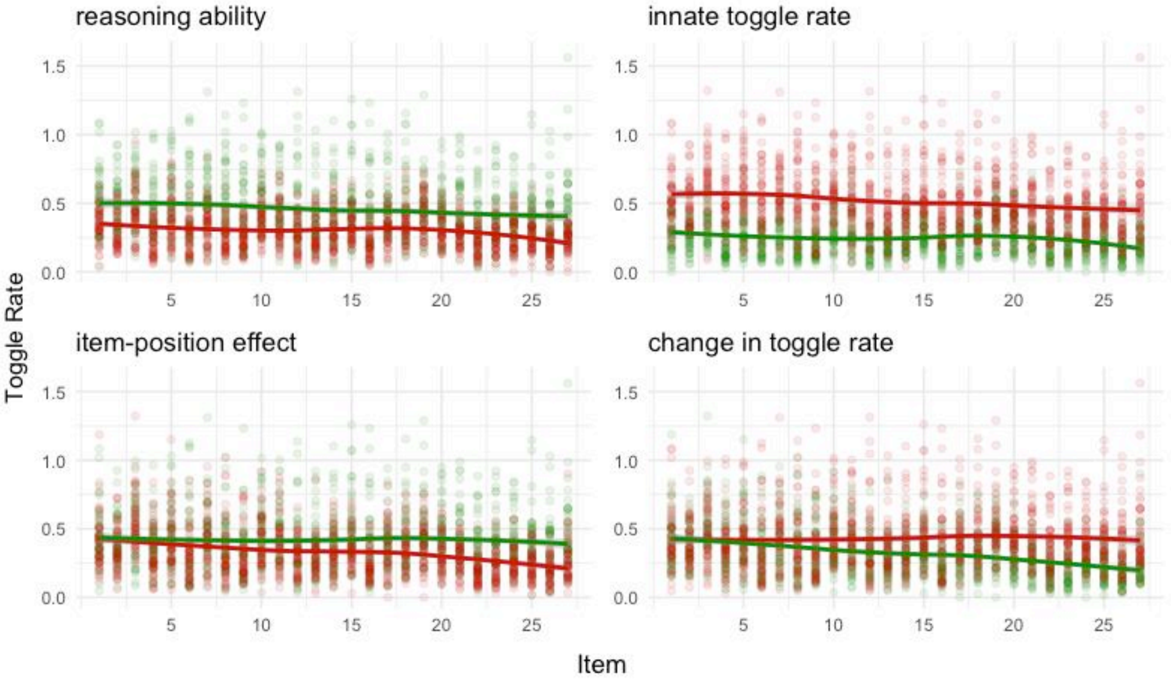


Figure 20

Plots of the 50 highest (red) and 50 lowest (green) factor scores on the respective latent variable



For Figure 20, the factor scores of the model illustrated in Figure 19 were extracted and the trend lines for the 50 highest and the 50 lowest factor scores of Toggle Rate throughout the test were depicted separately for the respective latent variables. Figure 20 is an adaption of Figure 9 in von Gugelberg and Troche (*in preparation*) given on page 57 in the appendix.

The top left corner of Figure 20 displays the Toggle Rate for participants with high (red) and low (green) factor scores on the latent variable for reasoning ability. Hence participants with high factor scores in reasoning ability exhibit lower Toggle Rate throughout the test compared to participants with low factor scores in reasoning ability. This indicates that low reasoning ability coincides with response elimination and higher reasoning ability with constructive matching, which would be in line with results of Vigneau et al. (2006).

The top right corner in Figure 20 shows that participants with high (red) factor scores in basic Toggle Rate also exhibit an overall higher Toggle Rate. Hence the latent variable for innate Toggle Rate reflects overall Toggle Rate and participants with low (green) factor scores in innate Toggle Rate most likely engage in constructive matching, and participants with high values (red) in response elimination.

Low (green) factor scores on the item-position effect (bottom left) seems to coincide with a steady Toggle Rate, not noticeably changing throughout the test. This could indicate that participants do not change their strategy but remain with the strategy applied during the first few items. Participants exhibiting high factor scores (red) on the latent variable depicting the item-position effect, seem to reduce their Toggle Rate. This means, that these participants adapt their strategy throughout the test, and gradually move towards more constructive matching (i.e., lower Toggle Rate).

For the latent variable depicting the change in Toggle Rate (bottom right) high factor scores (red) coincide with a stable Toggle Rate. Low values (green) on the other hand, depict a decrease in Toggle Rate. This indicates that the latent variable depicting change in Toggle Rate, captures the negative (downward) change in Toggle Rate, i.e., a gradual shift towards more constructive matching and less response elimination for participants with low factor scores.

Theses descriptive results also translate nicely to the correlations depicted in Figure 19. A high value on the latent variable depicting the item-position effect (bottom left, red line) is related to a low value in change of Toggle Rate (bottom right, green line). Also, a high value on the latent variable depicting innate Toggle Rate (top right, red line) coincides with a low value in reasoning ability (top right, green line). Hence, the latent variable for reasoning ability showed a large negative correlation with innate Toggle Rate, but not the change of Toggle Rate. And the item-position effect showed a large negative correlation with the change in Toggle Rate, but not innate Toggle Rate.

Study 3: Rule disruption and the Learning Hypothesis

Participants in both conditions achieved similar scores in the first 18 items. Details can be taken from Table 4 and Figure 21. For both conditions three initial models were fit to the data. Model descriptions and fixations can be taken from Table 1. Model fit for all calculated models can be taken from Tables 3 and 4 in von Gugelberg et al. (2025)

Table 4

Duplicate of Table 2 in von Gugelberg et al., (2025)

Descriptive test statistics for the figural analogies test (TFA)

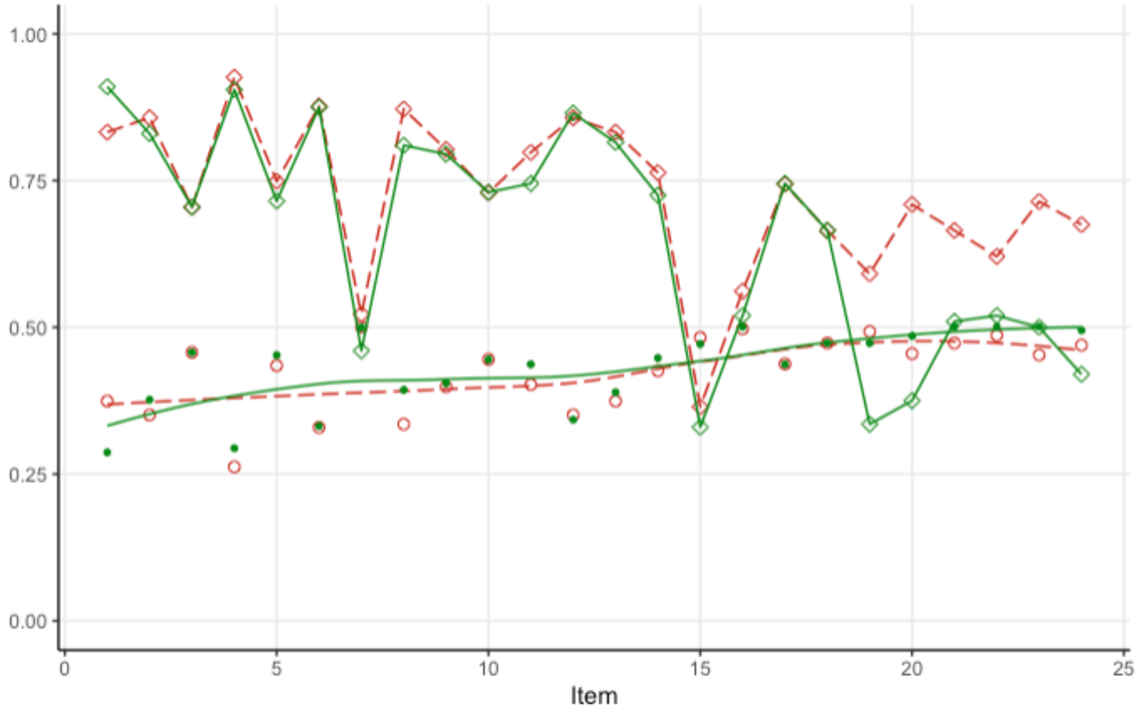
Condition	Items	Mean	SD	Min	Max	Skewness	Kurtosis	Cronbach's α
continuous rule								
(*n* = 203)	1-18	13.46	3.58	3	18	-0.81	0.04	.81
	1-24	17.43	5.05	3	24	-0.75	-0.29	.86
discontinuous rule								
(*n* = 200)	1-18	13.15	3.56	2	18	-0.90	0.42	.80
	1-24	15.81	4.73	2	24	-0.51	-0.08	.83

Note. For each condition the first 18 items (1-18) and the full set of 24 items (1-24) with their respective descriptive statistics are presented.

For the continuous rule condition, the bifactor model with factor loadings increasing linearly for the latent variable depicting the item-position effect showed the best fit (see Figure 22 Panel A for illustration). Hence the TFA created, successfully elicited an item-position effect. For the discontinuous rule condition none of the three described models in Table 1 yielded an acceptable fit. Therefore, the manipulation of rules did not allow for configural invariance, which lets us conclude that the manipulation of rules was successful in disrupting test taking behavior.

Figure 21

Standard deviation and accuracy for every item in both conditions.



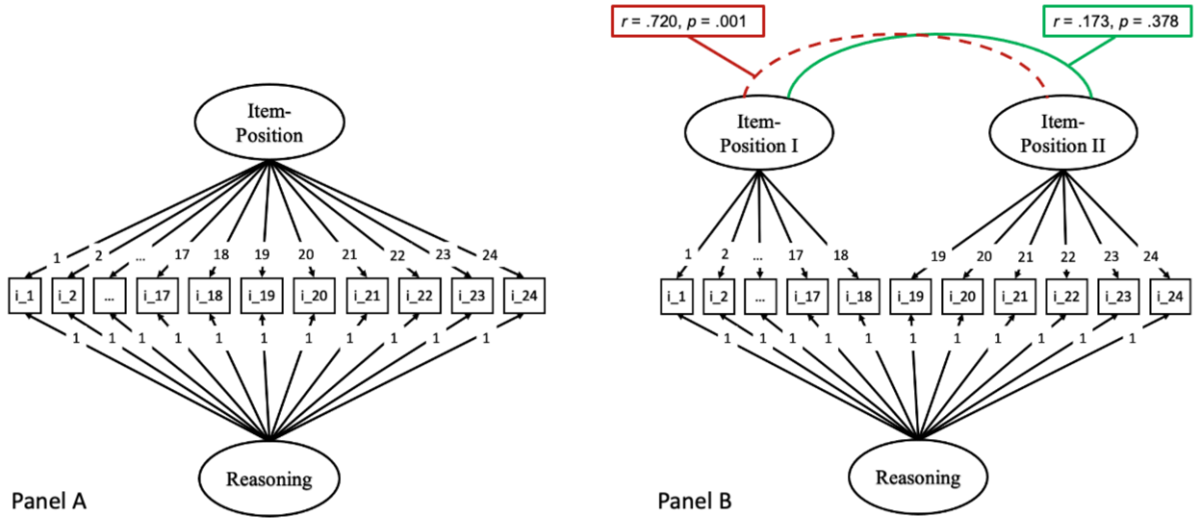
Note. Values of the discontinuous rule condition are depicted with solid lines, for the continuous rule condition with dashed lines. Item accuracy is depicted with squares, the standard deviation with filled and empty circles in the continuous and in the discontinuous rule condition, respectively. Smooth lines indicate trend lines for the standard deviations of the two conditions.

To describe the data of the discontinuous rule condition well, a third latent variable was needed (see Figure 22, Panel B for illustration). When the model included one latent variable depicting reasoning, one for the item-position effect in the first 18 items, and one for an additional item-position effect for the last six items, data description improved. Whether the correlation between the two latent variables depicting the item-position effects was estimated or set to zero hardly changed model fit. This makes sense, since the estimated correlation was small and insignificant (see solid green line in Figure 22).

When the same model was fit to the continuous rule condition, the estimated correlation between the two latent variables depicting the item-position effects was large and significant (see dashed red line in Figure 22). Fit of said model was just a tad better than the bifactor solution with two latent variables. The large correlation between the two item-position effects indicates substantial overlap between the two latent variables. Hence the more parsimonious solution with one latent variable depicting the item-position effect across all 24 items most likely provides the best data description.

Figure 22

Bifactor models with two or three latent variables



Note. Panel A is a simplified illustration of a bifactor model with two latent variables. Panel B displays a bifactor model with three latent variables. Connecting lines between the two latent variables inform about the correlation between the two latent variables if estimated freely and not set to zero. Correlation of the discontinuous rule condition is depicted with a solid line, for the continuous rule condition with a dashed line.

Several alternative explanations for the different response behavior between conditions were explored. This includes, but is not limited to item difficulty, mean response latencies, and number of rules per item. More detailed descriptions about these alternative models can be found in the Supplementary Materials of von Gugelberg et al., (2025) and a short description in the article itself under *3.1 Supplementary Analyses*.

One alternative explanation could not be ruled out to a satisfying degree. An anonymous reviewer pointed out the work of Nagy et al. (2023) and the possibility of rapid guessing or disengagement during test taking. The reviewer highlighted the possibility, that the introduction of new rules could have influenced the number of participants applying rapid guessing or disengaging while completing the last six items. While there was no difference between conditions in self-reported test taking diligence (on average the answer was around 7.5 on a 9-point Likert scale for both conditions), self-reports are far from objective and are not enough to rule out rapid guessing or disengagement. A simple analysis of RT is also moot, since according to Nagy et al. (2023) rapid guessing would entail very short RTs and disengagement rather long RTs, possibly canceling each other out. While Nagy et al. (2023) propose to include measures of test taking persistence, with randomized item order to account for both rapid guessing and disengagement, this is something we cannot do in hindsight, leaving these questions unanswered for the moment.

Discussion

First, I will briefly reiterate the discussed points for each study with a focus on the relevant findings for this dissertation. This is followed by the Integrative Discussion, where beyond what was discussed prior in the studies I will highlight the specific contributions to the field, address what questions remain unanswered and propose future avenues for research.

Study 1: Attentional control and the Learning Hypothesis

The dual mechanism of control theory distinguishes between proactive and reactive attentional control. Gray et al. (2003) and Burgess and Braver (2010) concluded from their results that higher fluid intelligence enables or facilitates the use of proactive control. We proposed the opposite, where the use of proactive control results in higher scores on a reasoning ability measure possibly due to a more pronounced item-position effect.

Within the RT pattern of the AXCPT one group with typical pattern for proactive control (Group A), one for reactive control (Group C) and another somewhere in between (Group B) were identified. While the three groups did differ in their reasoning ability scores, these differences were no longer significant when the item-position effect was accounted for. Instead, as displayed in Figure 15 (or Table 7 in von Gugelberg et al., 2021), Group A and C had significantly different factor scores in the latent variable depicting the item-position effect.

This difference in factors scores indicates that Group A had a more pronounced item-position effect compared to Group C. It therefore seems that individuals engaging in proactive control (Group A) benefit more from solving previous items when they solve later items in a reasoning test compared to individuals engaging in reactive control (Group C). Hence, proactive control has individuals use context information, and knowledge gained from solving previous items (i.e., rules), to solve later items. This supports the idea that rule learning underlies the item-position effect. As Braver (2012) states, proactive control allows for the maintenance of context information and guides attention ideally for the task at hand. In a reasoning ability measure, the task at hand is identifying and solving rules. Since rules are repeatedly used, a focus on the rules applied, maintaining their relevant indicators to guide attention, is very beneficial for the task of solving the next item. This benefit, we assume would also grow with each item solved, since rule knowledge can be accumulated.

While the conclusion seems plausible from the current results, the study design does not allow for casual interpretation of effects. Our results do provide an alternative explanation for the relation between fluid intelligence and cognitive control suggested by Burgess and Braver (2010). Yet further research is needed to determine direction of effect in a satisfactory manner.

Future research should aim for a more balanced distribution between groups engaging in different mechanism of cognitive control. Group C in our sample was rather small, and large effect sizes were not significant. Also, with the AXCPT we were successful in distinguishing different mechanisms of control, but data indicated three groups and not the presumed two groups in the dual mechanism of cognitive control theory. This calls for further exploration of the proposed dualism of cognitive control and its best practices of measurement.

Study 2: Strategy and the Learning Hypothesis

It has been frequently observed that individuals rely on two distinct strategies when solving common reasoning ability measures (e.g., Bethel-Fox et al., 1984). It is unclear what influences strategy selection. Several studies found a positive correlation between constructive matching and fluid intelligence (e.g., Jastrzebski et al., 2018), others detected a shift in strategy throughout test completion (e.g., Gonthier & Roulin 2020). Using a fixed-link model to analyze the Toggle Rate, we found both. A bifactor model with one latent variable depicting an innate difference in Toggle Rate accompanied by a second latent variable depicting a change in Toggle Rate outperformed a one factor solution. For the APM score data a bifactor solution also outperformed a one factor solution, enabling us to test the main hypothesis.

Under the premise that rule learning underlies the item-position effect I assumed that constructive matching would foster rule learning and hence would be associated with the item-position effect. Results indicated no relation between the latent variable of the item-position effect and innate Toggle Rate, but a large negative correlation with the change in Toggle Rate. This relation indicates that individuals with a pronounced item-position effect were more likely to gradually engage in more constructive matching (i.e., lower Toggle Rate) as the test progressed. This does support the assumption of rule learning underlying the item-position effect. When completing a reasoning ability measure, more opportunities arise to learn the given rules since rules are used repeatedly. Hence, the individual gradually accumulates rule knowledge making constructive matching a more expedient strategy with every rule repetition. Individuals with a very faint item-position effect, do not seem to exhibit a change in strategy. This indicates that these individuals stick with the strategy applied during the very first items, without adapting their strategy as the test progresses. This could be due to them not applying the knowledge gained from prior items to solve later items. In terms of the learning hypothesis, this would mean, that they do not accumulate knowledge about the rules used, or at the very least, they do not seem to apply any of the gained knowledge to subsequent items.

While current results seem to support the learning hypothesis, the detected change in Toggle Rate towards more constructive matching is somewhat the opposite of what Gonthier and Roulin (2020) found in their data with questionnaires. Their data indicated a shift towards response elimination around the 25th item and a decrease in item latencies around the 30th item. Authors believe this decrease most likely to be due to participants disengaging as items became to taxing. While item latencies also increased upon the 30th item in the current sample, there was no noticeable drop of item latencies thereafter. Item latencies remained rather constant after the 30th item, but it must be mentioned, that due to the implemented time limit in the current study, these item latencies are biased, since they lack individuals working through the APM at a slower pace.

These qualitative differences between samples, the operationalization of strategy with questionnaires not accounting for item latencies, and the different analysis approach may explain the difference in results, but additional studies are necessary.

For example, studies of an explorative nature could provide more insight, since recent research has identified a third strategy (Jarosz, et al., 2019). Using think-out-loud protocols the isolate-eliminate strategy was discovered. With this strategy participants eliminate bad lures, to increase their likelihood of selecting the correct answer from the remaining response alternatives. Li et al. (2022) ran an LPA on their questionnaire data and also found a third strategy. These individuals showed high scores for constructive matching AND response elimination. The three groups identified differed in reasoning ability scores but not in Toggle Rate. Their results imply that Toggle Rate might not be able to detect the third strategy identified.

The number of strategies, and what properties initiate a shift in strategy throughout a test is still unclear. Further research on the topic is not only relevant under the premise of the learning hypothesis but also since, in line with other research (e.g., Jastrzebski et al., 2018) the innate difference in Toggle Rate was strongly related to reasoning ability. A lower Toggle Rate indicating more constructive matching, was related to higher values on the latent variable depicting reasoning ability. Constructive matching seems to coincide with higher performance on reasoning ability measures. Interestingly Gonthier and Thomassin (2015) found that participants can be manipulated to use more constructive matching. Hayes et al. (2015) showed that in a test – retest setting one third of the explained variance in score gains were due to strategy. This information is especially interesting since differences in strategy use can already be observed in young children (Starr et al., 2018). Further underlining the importance of fully understanding the impact of strategy use during test taking and what can influence strategy selection.

Study 3: Rule disruption and the Learning Hypothesis

To directly test, whether rule learning underlies the item-potion effect two reasoning ability measures were created. In the continuous rule condition, the same rule was used for all 24 items and a typical item-position effect was observed (see Figure 22, Panel A). In the discontinuous rule condition, the first 18 items were identical to the continuous rule condition, followed by six items with different rules. The item-position effect was disrupted in the discontinuous rule condition since a three factor solution (see Figure 22, Panel B) was necessary to describe the factorial structure adequately.

While the experiment allowed us to rule out several alternative explanations (e.g., item difficulty, response times, etc.,) one alternative explanation remains. Future study designs must account for the possibly of rapid guessing and disengagement as in Nagy et al. (2023) and implement the adequate controls.

Additionally, studies should investigate whether the possible rule learning during test taking is of an explicit or implicit nature. On the one hand, two previous studies (Ren et al., 2014; Schweizer et al., 2021) where complex rule learning was strongly correlated with the item-position effect, participants were explicitly instructed in the complex rule learning task. This suggests explicit rule learning to be depicted by the item-position effect. On the other hand, the item-position effect is often unrelated to the reasoning ability in measurement models when the correlation is freely estimated (Schweizer et al., 2021). This was also the case in the continuous rule condition and has also been reported for measures of implicit learning (e.g., Danner et al., 2017; Kalra et al., 2019), not allowing for a definitive conclusion on the matter.

With the current results it is also unclear how many rule repetitions are needed to elicit an item-position effect. Since the factor loadings for the second item-position effect in the discontinuous rule conditions did not show an obvious increase when estimated freely, six items most likely were not enough. Future research should investigate this and include external variables to better determine the nature of the item-position effect and the possibly underling rule learning.

With the theory driven creation of items for the two version of the TFA, new avenues to explore the item-position effect have been made accessible. We successfully created an experimental design that can elicit and disrupt an item-position effect enabling future research to combine this experimental approach with correlational analyses.

Integrative Discussion

First, I will elaborate on the presented studies in this dissertation and propose several directions for worthwhile future research *on the item-position effect and the learning hypothesis* by including available information on the item-position effect reliant on the fixed-links approach. Then, considering a somewhat broader definition of the item-position effect follows a section *on culture and the item-position effect*. To understand a psychological phenomenon in a comprehensive manner, it is my opinion, that the proposed theories and explanations must be evaluated in different cultural contexts. Only then can conclusions be drawn whether a phenomenon is universal to all humankind. Thereafter, I will elaborate *on intelligence models and the item-position effect*. Here the goal is to open a discussion on what role the underlying factors of the item-position effect could play in the construct of human intelligence and to discover new and important directions for research concerning the item-position effect and the possibly related rule learning. The Integrative Discussion is concluded by a brief discussion *on methods and the item-position effect*, presenting additional methodological approaches to advance the field.

On the Item-Position Effect and the Learning Hypothesis

I took a closer look at the learning hypothesis proposed by Ren et al. (2014). During this undertaking we discovered a possible alternative direction of effects in regard to reasoning ability and attentional control due to a possible bias in favor of the underlying rules in an item (von Gugelberg et al., 2021). We were able to reveal a strong negative relation between the item-position effect and the change in Toggle Rate, pointing towards a shift in strategy that possibly fosters rule learning (von Gugelberg & Troche, *in preparation*). And we successfully created a reasoning ability measure, wherein the simple manipulation of rules disrupted the item-position effect (von Gugelberg et al., 2025).

Revisiting the latter results, it is still unclear, why the two item-position effects in the discontinuous rule condition were not related. It is possible, that six items with rule repetitions were simply not enough to elicit an item-position effect, since six items did not provide enough learning opportunities. It would be interesting to see, whether the same experimental design of the discontinuous rule condition but with additional items based on the two new rules would elicit an item-position effect. Such a study design could address how many items are truly necessary to elicit an item-position effect. From the discontinuous rule condition, one can conclude that 18 items seem to be enough, but six items were not, since factor loadings did not really depict an increase.

Consulting other work on the item-position effect the inexistent correlation between the two item-position effects in the discontinuous rule condition might not have been due just to a lack of items for the second item-position effect. In Schweizer, Reiss et al. (2012) no correlation was found between the item-position effects of a version of the APM (Raven et al., 1998) and the Horn reasoning scale (1983). Neither was one detected between another Figural Matrices Test (Kyllonen et al., 2019) and the first 18 items of the TFA (von Gugelberg & Troche, 2022a). Similarly, the item-position effects did not correlate across different time points (Wang et al., 2020), yet the item-position effects between the different subtest of the CFT did (Troche et al., 2016). This raises the general question whether the item-position effect truly captures something as essential as a rule learning ability, when it does not share any variance across different measurements of reasoning ability. The item-position effect might be strongly biased by yet unknown entities, making any amount of shared variance vanish.

To further investigate this, one could set up a somewhat more elaborate experimental design using the figural analogies of von Gugelberg et al. (2025). It would allow for identical test setting and stimulus material giving little wiggle room for unknown confounding variables. For the new experiment again two sets of TFA's could be created. One set would contain the first 18 items used in both conditions of von Gugelberg et al. (2025) and the 18 items for the second set would be created with the two other rules (line removal, mirroring). One group of participants would then start with the original set of 18 items followed by the new set, the other group vice versa. Theoretically the item-position effect should be disrupted in both groups, and if it is some sort of rule learning ability, the item-position effects within both groups should correlate with one another at least to some degree.

Without having more specific experiments on the topic, drawing further conclusion is difficult and maybe a broader look at all the conducted research on the item-position effect relying on fixed-links modelling (Schweizer, 2006) might provide some insights or different angles to approach the questions: What is the item-position effect? What effect can be disrupted by a simple change of rules in a reasoning ability measure (von Gugelberg et al., 2025)? Is it truly rule learning that hides behind the item-position effect?

Under the premise of the learning hypothesis, two studies could show, that the item-position effect is correlated with complex learning tasks (Ren et al., 2014; Schweizer et al., 2021). The item-position effect also served as a better predictor for students' academic performance presumably due the item-position effect depicting a type of learning ability (Ren et al., 2015). Schweizer et al. (2020) found a relation of the item-position effect to rule acquisition and sustained attention. Additionally, studies presented in this dissertation under the premise of the learning hypothesis could link the item-position effect to proactive control (von Gugelberg et al., 2021), a shift in strategy towards more constructive matching (von Gugelberg & Troche, *in preparation*) and its need for continuous rule presentation (von Gugelberg et al., 2025).

Outside the premise of the learning hypothesis findings on the item-position effect with fixed-links modeling include but are not limited to an early study by Lozano (2015) exhibiting a correlation between the item-position effect and impulsivity, while Ren, Gong et al. (2017) found no such correlation (neither did Krampen et al., 2020) but rather a strong positive correlation with executive attention. Considering further executive functions, the item-position effect was associated with updating and shifting, but not inhibition, when analyzing APM scores (Ren, Schweizer et al., 2017).

In a somewhat different line of research Sun et al. (2019) found no item-position effect in 7-8 year olds, but for 12-13 year olds an effect emerged in the data, and the correlation of reasoning ability and working memory considerably increased when the item-position effect was accounted for in the model. Another study found an item-position effect in children as young as 10 years old (Wang et al., 2020).

This research on the item-position effect in young children could provide a link to the explored connection of strategy use and the item-position effect in von Gugelberg and Troche (*in preparation*). For example, results of Thibaut and French (2016) indicate that 5- and 8-year-olds alternate more often between the figures in the given analogy and possible response alternatives than adolescents. While the authors do not directly analyze this, children's behavior is more in line with the response elimination strategy and the tested adolescents and adults spend noticeably more time observing the initially given figures in the analogy, mimicking the constructive matching strategy. Authors point towards the possibility of children having difficulties in inhibiting the main goal of the task (i.e., finding a solution among the response alternatives).

Starr et al. (2018) found the same pattern of results in 6-year-olds, but also identified large individual differences. Their results indicated that children spending more time analyzing the given analogy before consulting the response alternatives (just as is characteristic for constructive matching) outperformed the children spending most of their time switching between the given analogy and the response alternatives. This indicates, that similarly as for example Jarosz et al. (2019) found a positive correlation between constructive matching and reasoning ability in adults, Starr et al. (2018) described the same phenomenon in children as young as 6 years old.

While children must at least have reached the age of 10 for an item-position effect to emerge (Wang et al., 2020), the established difference in strategy use found in adults (e.g., Jarosz et al., 2019), seems to already be present in younger children (e.g., Starr et al., 2018). If one assumes from the results of von Gugelberg and Troche (*in preparation*) that the item-position effect is related to a change in strategy towards more constructive matching during test completion, such a change should emerge in children above the age of 10 (Wang et al., 2020), but not any younger (Sun et al., 2019). This could provide another interesting avenue to explore.

Consulting literature a further similarity between children and adults concerning strategy use in reasoning ability measures can be found. For example, Glady et al. (2017) found, that young children (4;7 -6;4 years old) can be successfully manipulated to focus more on the given analogy (as is common for the constructive matching strategy), which lead to improved task performance. Gonthier and Thomassin (2015) found this to be true for adults also. Interestingly Glady et al. (2017) also found if there was a very salient, but irrelevant distractor present in the analogy, the strategy manipulation did not work in favor of task performance. Authors conclude that possibly children had more difficulties dealing with the interference caused by the salient distractor.

It would be interesting to investigate whether there are individual differences in the ability to deal with such interference in young children but also adults. This would be an especially interesting undertaking since the proactive mechanism of cognitive control, coined by Braver (2012) in the dual mechanism of attentional control theory seems to be related with the ability to deal with interferences. Results of Gray et al. (2003) suggest that the stronger engagement in proactive control, by individuals with high reasoning abilities lead to a better performance when interference was high, compared to individuals with lower reasoning abilities engaging in reactive control.

Individuals engaging in proactive control also showed the most pronounced item-position effect (von Gugelberg et al., 2021). Regarding the disruption of the item-position effect in von Gugelberg et al., (2025), the sudden change in rules, created salient but irrelevant distractors. Both the dot and the trapezium (see Figure 6) were essential in the first 18 items to solve an item correctly, since they were manipulated with the movement rule (Figure 7, Panel A). When the rules suddenly changed in the discontinuous rule condition, they were no longer relevant, since straight lines were subtracted, and the figure was mirrored (Figure 7, Panel B). Nonetheless, if an individual successfully learned the movement rule while completing the first 18 items, their attention now is biased towards the dot and the trapezium, making them very salient distractors.

From Grey et al. (2003) one would assume that individuals predominantly engaging in proactive control would be better at dealing with this interference caused by the salient distractors. Additionally, from the results in von Gugelberg et al. (2021) one would assume, that individuals predominantly engaging in proactive control, would also exhibit the most pronounced item-position effect. This would indicate, that in the study of von Gugelberg et al. (2025), the participants with the most pronounced item-position effect in the first 18 items of the discontinuous rule condition, should be able to handle the distractors (dot & trapezium) very well and experience the overall smallest decline in performance for the last six items were the new rules were implemented. Of course this is highly speculative, and in dire need of specific studies addressing the topic.

One possibility to investigate this could include measuring eye movements while participants solve the TFA version used for the discontinuous rule condition (von Gugelberg et al., 2025). By setting the interest areas directly on the dot and trapezium of each item, it is possible to analyze the frequency of fixations on them. While within the first 18 items the dot and trapezium provide necessary information, they no longer do so after the rules behind item creation have changed. Thus, making them very salient distractors. If the item-position effect depicts some sort of rule learning, participants with a very pronounced item-position effect should in theory show more fixations on the rule relevant elements in an item than participants with a very faint item-position effect. Within the first 18 items that would be of course the dot and trapezium, after the rules were changed and the dot and trapezium are no longer relevant, individuals with a pronounced item-position effect should be able to deal well with the salient distractors. This would translate to fewer fixations on the salient distractors. Individuals with a mediocre or faint item-position effect would overall show fewer fixations on the rule relevant elements within the first 18 items, but also would be less effective in handling the interference the dot and trapezium create, after the rules underlying the items have changed.

Such a study design could also provide further detail on strategy use, interference and rule learning. Taking a closer look at the results of von Gugelberg and Troche (*in preparation*), it is not the initial strategy applied by the individual that is related to the item-position effect. It is the adaptive behavior during test taking. In the case of the APM it was an adaptive behavior towards more constructive matching. Possibly the proactive mechanism of control allowed the individuals to bias their attention towards identifying rules already learnt in prior items for the current item. Similarly, they could handle interference through irrelevant information in a given item well. And with the rule repetitions in the APM and increasing item difficulty, the best way to adapt their response behavior was a shift towards more constructive matching. It might be possible, that with different test properties the adaptive behavior during test taking could take on a different form.

Raden and Jarosz (2020) found that test properties such as for example the ambiguity of an item can influence participants strategy behavior during test completion. An ambiguous item theoretically has more than one correct answer, and the participant must thus consult the presented response alternatives to decide on the correct solution. If ambiguous items are presented more frequently, participants shifted towards a strategy with more focus on the response alternatives.

Taking the results from Raden and Jarosz (2020) into account, it would be interesting to see whether similarly to von Gugelberg and Troche (*in preparation*) a change in strategy can be detected in a test with more ambiguous items, whether this change would be towards more response elimination, and whether this change would be related to the item-position effect. If results would support this assumption, it might be reasonable to assume a new theory behind the item-position effect. It is not the ad hoc rule learning, but the adaptive behavior during test taking that is captured by the item-position effect. Yes, adaptive behavior during test taking would be related to ad hoc rule learning, but it also would include the ability to deal with interferences, goal maintenance, rule retrieval, and it would vary a lot between different tasks to what degree the different abilities are needed for the ideal adaptive behavior.

If it were the adaptive behavior during test taking that underlies the item-position effect it could explain why the different item-position effects are most often not related to each other. Different tests use different abilities, and just because some individuals are very good at one thing, it does not mean they are all equally good at other things necessary for adaptive behavior during test taking. This does not contradict the leaning hypothesis, but rather is the use of a broader terminology to describe a phenomenon that possibly underlies the item-position effect.

On Culture and the Item-Position Effect

In my opinion, it is important to evaluate whether a psychological phenomenon is universally applicable in order to understand it to a satisfying degree. Hence it is paramount to investigate whether the individual differences in test taking behaviour occur to a similar degree in different cultural settings. Unfortunately, there is a cultural bias in psychological research since a vast number of psychological measures were developed and validated in WEIRD[2] cultures (e.g., Nielsen et al., 2017) and still many cultural groups are underrepresented (Krys et al., 2024). Nonetheless, a short excursion into Sternberg's (2019) theory of adaptive intelligence seems a fruitful endeavour as a first step in addressing this topic. Especially if one uses the somewhat broader definition of adaptive test taking behaviour possibly underling the item-position effect.

Sternberg (2019) defines intelligence as an "adaption to the environment" which in broad terms also translates to the definitions of general intelligence. Yet he draws a clear distinction between the two. He postulates, that what today is being measured with common intelligence tests, including reasoning ability measures, is simply one specific instance of how adaptive intelligence can manifest itself. He states: "General intelligence, as measured by Western psychometric tests and cognitive tasks, is not a necessary condition for adaptive intelligence across cultures" (p.3, Sternberg, 2019). Thus, what is commonly measured relates to successful adaptive behaviour found in Western or also WEIRD cultures, but by no means also adequately measures successful adaptive behaviour in other cultures. Sternberg (2007) explains in detail, what intelligence can mean in different cultures, and how different skills or abilities to adapt can predict a prosperous life in different cultures[3]. One form of adaptive behaviour in the theory of adaptive intelligence refers to the ability to adapt one's

[2] WEIRD, is an acronym for people from a Western, Educated, Industrialized, Rich and Democratic populations.

[3] Sternberg (2019) calls for a definition of adaptive intelligence where the term "adaption" is used in a broad manner. That is, not only adapting oneself (and the behaviour) to deal with the environment, but also the ability to adapt the environment to fit oneself and finding or creating new environments as needed.

own behaviour to the environment. When solving a reasoning ability measure, its test properties would represent the environment, and ideally adapting one's own solving behaviour to give a correct answer, could be seen as the adaptive intelligent behaviour.

Applying Sternberg's (2021) theory to the item-position effect in common Western reasoning ability measures, in a simplified manner could indicate, that the latent variable representing reasoning ability would refer to a culture specific intelligence. With such measures created and validated in WEIRD cultures, they measure a problem-solving behaviour that is very relevant in WEIRD cultures, but possibly not in other cultures. The item-position effect on the other hand, could either represent culture specific adaptive behaviour, or a culture independent adaptive behaviour.

As a follow-up to Study 2 (von Gugelberg & Troche, *in preparation*) it would be worthwhile to investigate whether the item-position effect shows similar relations to strategy use and change in strategy in non-WEIRD cultures. Since a simple task such as finding Waldo in a "Where is Waldo?" book indicated different search patterns / scan paths between two cultures when analysing eye movements (Lüthold et al., 2018), scan paths (i.e., strategy) could also differ between cultures when solving reasoning ability measures. Further it seems important to investigate whether the additional strategy found in Chinese participants when solving reasoning ability measures in Liu et al. (2023) and Li et al. (2022) truly translate to findings from for example Jarosz et al. (2019) in a sample of US students with think-out-loud protocols.

The question whether the item-position effect captures culture specific adaptive behaviour, or culture independent adaptive behaviour also shines a new light on the findings of Ren et al. (2015) in a sample of Chinese children. Their results indicated that the item-position effect was a better predictor for academic achievement than reasoning ability itself. Therefore, future studies should investigate whether a reasoning ability measure is an inferior predictor for school performance also in non-WEIRD / Western cultures when the item-position effect is accounted for.

Additionally, it is still unclear, whether it is the same adaptive behaviour during test taking captured by the item-position effect even with it emerging in European (e.g., Ren, Wang et al., 2014) and Chinese (e.g., Sun et al., 2019) samples. Since the experimental design of Study 3 (von Gugelberg et al., 2025) was implemented online, a direct replication in for example a Chinese sample seems highly feasible. If the same results emerge, one could at least conclude that the item-position effect can be disrupted by a sudden change of rules, independent of culture. Alas it would still be unknown, whether the idiosyncratic behaviour of individuals is similar across cultures.

On Intelligence Models and the Item-Position Effect

For a thorough investigation of the item-position effect, attempts to imbed it in common intelligence models also must be made. It allows to address, whether the ad hoc rule learning or adaptive test taking behaviour possibly captured by the item-position effect is a part of common intelligence models or must be seen as an entity separate from intelligence in common models of intelligence today.

When relying on the assumption that the variance in test taking behavior captured by the latent variable depicting the item-position effect is the ability to ideally adapt one's test taking behavior according to the given circumstances, different measures likely require different abilities to a different degree. Therefore, it seems that some abilities would be very test specific, and others could be more general, required by more than one task. The knowledgeable reader familiar with intelligence theories would immediately be reminded of Spearman's two-factor theory of intelligence postulated in 1927.

Spearman assumed one general factor for intelligence (g) that was to be correlated with specific abilities (s) measured with different tests. His theory stemmed from observations of correlation matrices and later from his pioneering work on factor analyses (1927). From performances of participants in various cognitive tasks Spearman concluded that some abilities could be clustered together and measured specific abilities (s), but "all branches of intellectual activity have in common one fundamental function" that he named *g* (Spearman, 1904, p. 284).

Alas, the structure of the two-factor theory does not translate to the observations made about the item-position effect. The fact that different item-position effects often do not share any variance (e.g., von Gugelberg & Troche, 2022a), means that they cannot co-vary, and thus not be captured by one common factor. Every item-position effect captures unique test specific variance, which to my knowledge was only found to correlate with each other in one instance (i.e., Troche et al., 2016). It seems more feasible to place the item-position effect within the two-factor theory of Spearman rather than applying the structure to the occurrence of the item-position effect itself. Meaning, that the item-position effect, could, in theory, be capturing a test specific ability (s) that cannot be explained by the common *g* factor. This could also explain the non-significant correlation between the item-position effect and reasoning (von Gugelberg et al., 2025), since reasoning would mainly be explained by the common *g* factor (Spearman, 1927).

Next to Sternberg's idea of adaptive intelligence (2021) and Spearman's common *g* factor (1927), the Cattell-Horn-Carroll (CHC) theory is a widely recognized theory on the underlying cognitive abilities of intelligence (e.g., Ortiz, 2015) and it would seem negligent not to mention it here. To allow for an adequate description of the CHC theory one must revisit some of the history of intelligence research. Things most likely started when Cattell (1941) despite having Spearman as his mentor, assumed intelligence not to be a single constant factor, but more likely to consists of multiple abilities. He paid attention to the possibility that abilities could develop, peak, or decline across the life span, which ultimately led Cattell (1943) to the Gf-Gc theory. The Gf-Gc theory depicts the distinction of fluid (Gf) and crystalline intelligence (Gc). Where crystalline intelligence broadly refers to knowledge attained throughout life and fluid intelligence (Gf) the ability to solve new problems (Cattell, 1963) which peaks during young adulthood. A model still frequently relied on today (e.g., Flanagan et al., 2000, Kyllonen & Kell, 2017).

Horn (1965) later added more abilities to the theory, all the while opposing the idea of general intelligence factor (*g*). In 1966 the theory included nine secondary abilities identified through a factor analysis (Horn & Cattell, 1966). By 1991 the theory was referred to as the extended Gf-Gc theory and Horn described a total of 42 primary (or narrow) abilities relevant for the 9 secondary (or broad) abilities in the Gf-Gc theory. Among the 42 primary abilities listed, summarized by the nine secondary abilities, none directly referred to rule learning abilities or adaptive behavior. Hence not providing a direct home for adaptive (or rule learning) behavior during test taking possibly captured by the item-position effect.

The extended Gf-Gc theory was later endorsed by Carroll in his groundbreaking book on Human Cognitive Abilities (1993) and the story of the CHC theory continues. The book (Carroll, 1993) is not remarkable due to the mentioned endorsement, but rather due to its comprehensive review of intelligence research and the large-scale reanalysis of multiple datasets (approx. 460) relying on explorative (rather than confirmatory) factor analyses. From the analyzed data he came to postulate the Three-Stratum-Theory (3S). The first Stratum referred to about 65 primary abilities, the second to 8 secondary abilities, and the third to one general ability. While the first and second stratum are akin to the narrow and board abilities mentioned in the extended Gf-Gc theory, the assumption of a general factor in the third stratum is the most distinctive difference to the Gf-Gc theory.

In all the selected tests and listed primary abilities in Carroll's 3S theory (1993) I could not find any direct reference to the ability to adapt one's behavior during test taking, yet rule discovery is listed (p.211, 1993) as a subtype of the reasoning domain. Further, Carroll (1993) defines memory and learning processes as a second order (second stratum) cognitive ability (p. 634, see p. 302 for an overview of the domain). He states that ”Learning and memory are related because memory has to do with how the outcomes of learning are retained or forgotten" (p. 248, 1993). He highlights difficulties of separating the two (pp. 674-677) and simultaneously describes learning as a core ability of intelligence, citing the relevant work. While there is no specific mention of rule learning, this would possibly be where the variance captured by the item-position effect would co-vary most with other latent variables and would in itself warrant a dissertation on the topic of the item-position effect, learning and memory.

Moving on from Carroll's 3S theory (1993) to the CHC theory, McGrew (2023) a mentee of Carroll, states that the CHC theory is the results of an arranged marriage of convenience between the extended Gf-Gc model and the 3S theory. With Carroll (1993) endorsing the work of Cattell and Horn, his 3S theory was seen as an extension of the extended Gf-Gc model and thus was born the CHC-theory. The CHC-theory was further imbedded in intelligence research by McGrew's (2009) editorial in the journal of *Intelligence* and book chapters on the theory (e.g., Schneider & McGrew, 2012). To the introduction of the CHC theory Carroll writes "I am still not quite sure what caused or motivated it" (p.16, 2003). Where *it* refers to the name change from the Gf-Gc theory to the CHC theory in the technical manual of the revised Woodcock-Johnson cognitive test battery (Woodcock et al., 2001). McGrew (2023) writes that Carroll "was vexed that the term CHC theory had been so rapidly infused into the literature and, more importantly, incorrectly implied that he [Carroll], Cattel and Horn had agreed to a formal union of theories" (p. 29).

The most salient issue might be that Horn (e.g., 1991) was vehemently against the concept of a general factor, while Carroll (1993) clearly endorsed the idea. And this is a discussion still held today. A general factor postulated by Spearman (1904) and empirically evidence by for example, Thurstone's (1947) further implementations of rotations in the factor analytic approach point towards the phenomenon of the positive manifold. The positive manifold describes the fact, that different ability tests are positively correlated. Spearman (1927) provided the simple explanation, that a common factor (i.e., *g*), plays a crucial role in the performance of cognitive ability measures.

The Process Overlap Theory (POT) postulates an alternative explanation. The POT sees the positive manifold as an epiphenomenon of a variation of cognitive abilities sharing variance (Kovacs & Conway, 2016) and not one underlying general ability. It is an ongoing debate and for example, network analyses on the functional relationship of the positive manifold and different aspects of execute attention did not provide support for the POT (Troche et al., 2021).

It would be interesting to investigate the source of the positive manifold with the item-position effect accounted for in every subtest. A comprehensive investigation including a very large sample completing an exhaustive test battery (e.g., WISC-V, Wechsler, 2017), could reveal new information about the item-position effect itself and possibly provide worthwhile contributions to the discussion on the positive manifold.

On Methods and the Item-Position Effect

Network analyses are seen as a promising tool by van der Maas et al. (2017) to integrate processes of cognitive abilities and possibly allow for a unified theory of intelligence. I also believe that network models could provide interesting results on the item-position effect. Through the extraction of factor scores (as in Troche et al., 2021; von Gugelberg et al., 2021) from models including the item-position effect further information can be gained as to why item-position effects from different measures (von Gugelberg & Troche, 2022a) or different time points (Wang et al., 2020) are not associated, but in the one instance of the CFT are (Troche et al., 2016).

Latent Profile Analysis (LPA, as in von Gugelberg et al., 2021) also provides a unique opportunity to circumvent problems related to the analysis of individual differences based on reaction time data (see Draheim et al., 2024 for additional alternatives). Thus, allowing for more elaborate analyses on the item-position effect and what could possibly create individual differences therein. The two main problems with reaction time analysis regarding difference scores often leading to low reliability and an improper account of speed-accuracy interactions (e.g., Draheim et al., 2019) no longer overcomplicates interpretation of results. The LPA allows for a data driven exploration of patterns specific to certain groups of individuals in continuous data. The grouping variable can then be included in any further analysis without ever having to rely on score differences.

Similarly, the fixed-links approach used to depict the item-position effect (e.g., von Gugelberg & Troche, *in preparation*) can be used to account for different factors in experimental tasks, as demonstrated in Pahud et al. (2018) or by von Gugelberg and Troche

(2022b). This approach also can circumvent the difficulties presented by reaction time difference scores.

Looking beyond the fixed-links approach, a Multilevel Modelling (MLM) was demonstrated by Birney et al. (2017) to be a feasible approach to account for an item-position effect in the data, while exploring personality factors. Their results on the APM indicated that, with each progressing item (the next item in the sequence) the odds of solving the item correctly decreased for participants with high Neuroticism scores and increased for participants with low Neuroticisms scores. Birney et al. (2022) describe in detail how MLM can further our understanding of individual differences in the test taking process.

I believe research on the item-position effect and the possible underlying rule learning or adaptive test taking behavior will benefit from combining different methodological approaches included (but not limited to) and briefly introduced here. The successful implementation of the TFA in von Gugelberg et al., (2025) makes the combination of experimental and correlational analyses more accessible and is a further step in the direction of the much-needed reconnection of correlational and experimental approaches (Cronbach, 1957; Wilhelm & Kyllonen, 2021)

Conclusion

Overall, results are in favor of the learning hypothesis, albeit not unambiguously. Taking further studies on the item-position effect into account, we might have to accept the boarder definition of adaptive behavior during test taking for the item-position effect. This definition would include rule learning, but adaptive test taking behavior would also rely on other factors. Such adaptive behavior could be influenced by personality factors (e.g., Neuroticism in Birney et a., 2017), information retention (due to the interplay of learning and memory highlighted Carroll, 1993), goal maintenance (e.g., proactive mechanism of control in von Gugelberg et al., 2021), different scan path behavior (e.g., Lüthold et al., 2018), test properties (von Gugelberg et al., 2025), disengagement or rapid guessing (e.g., Nagy et al., 2023) and differences in adapting ones solving strategy during test taking (von Gugelberg & Troche, *in preparation*).

Future research on the item-position effect and rule learning or adaptive behavior during test taking can benefit from the successfully implemented TFA in von Gugelberg et al., (2025), by combining the experimental set up in the study with additional correlational analyses with external variables (e.g., pure information retention measures). The fixed-links method has also been made more accessible with the open-source software R (R Core Team,

2020) and the provided *bindabox* package (von Gugelberg, 2022). This will hopefully entice others to apply and explore the fixed-links approach also beyond the item-position effect. For example, in von Gugelberg and Troche (*in preparation*) we were able to successfully capture a change in the applied strategy occurring throughout a test, directly based on the observations made on eye movements. Thus, allowing to capture individual differences in test taking behavior.

With the undertaking of this dissertation and the three studies investigating the item-position effect regarding the learning hypothesis, new and relevant factors contributing to the item-position effect were identified. By consulting models of intelligence further relevant questions regarding the learning hypothesis and the item-position effect were discovered. Only with a more comprehensive understanding of what influences test taking behavior and its outcome, can one truly aim for complete test fairness across individuals and cultures. Therefore, this line of research is a cornerstone of psychological research, and further investigation is essential.

References

Akogul, S., & Erisoglu, M. (2017). An Approach for Determining the Number of Clusters in a Model-Based Cluster Analysis. *Entropy*, *19*(9), 452. https://doi.org/10.3390/e19090452

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*(3), 205–238. https://doi.org/10.1016/0160-2896(84)90009-6

Becker N. & Spinath, F. (2014). *Design a Matrix – Advanced.* Hogrefe.

Birney, D. P., Beckmann, J. F., Beckmann, N., & Double, K. S. (2017). Beyond the intellect: Complexity and learning trajectories in Raven's Progressive Matrices depend on self-regulatory processes and conative dispositions. *Intelligence*, *61*, 63–77. https://doi.org/10.1016/j.intell.2017.01.005

Birney, D. P., & Beckmann, J. F. (2022). Intelligence is cognitive flexibility: why multilevel models of within-individual processes are needed to realise this. *Journal of Intelligence*, *10*(3), 49. https://doi.org/10.3390/jintelligence10030049

Blum, D., & Holling, H. (2018). Automatic generation of figural analogies with the IMak package. *Frontiers in Psychology*, 1286. https://doi.org/10.3389/fpsyg.2018.01286

Blum, D., Holling, H., Galibert, M. S., & Forthmann, B. (2016). Task difficulty prediction of figural analogies. *Intelligence*, *56*, 72–81. https://doi.org/10.1016/j.intell.2016.03.001

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Braver, T. S., Barch, D. M., Keys, B. A., Carter, C. S., Cohen, J. D., Kaye, J. A., Janowsky, J. S., Taylor, S. F., Yesavage, J. A., Mumenthaler, M. S., Jagust, W. J., & Reed, B. R. (2001). Context processing in older adults: Evidence for a theory relating cognitive control to neurobiology in healthy aging. *Journal of Experimental Psychology: General*, *130*(4), 746–763. https://doi.org/10.1037/0096-3445.130.4.746

Burgess, G. C., & Braver, T. S. (2010). Neural mechanisms of interference control in working memory: Effects of interference expectancy and fluid intelligence. *PloS One*, *5*(9), e12861. https://doi.org/10.1371/journal.pone.0012861

Carlstedt, B., Gustafsson, J.-E., & Ullstadius, E. (2000). Item sequencing effects on the measurement of fluid intelligence. *Intelligence*, *28*(2), 145–160. https://doi.org/10.1016/S0160-2896(00)00034-9

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404. https://doi.org/10.1037/0033-295X.97.3.404

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. *The scientific study of general intelligence*, 5–21. https://doi.org/10.1016/B978-008043793-4/50036-

Cattell, R.B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin, 38,* 592.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin*, *40*(3), 153–193. https://doi.org/10.1037/h0059973

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1–22. https://doi.org/10.1037/h0046743

Cowan, N., Fristoe, N. M., Elliott, E. M., Brunner, R. P., & Saults, J. S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & cognition*, *34*, 1754-1768. https://doi.org/10.3758/BF03195936

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671–684. https://doi.org/10.1037/h0043943

Davey, G., De Lian, C., & Higgins, L. (2007). The university entrance examination system in China. *Journal of further and Higher Education*, *31*(4), 385-396. https://doi.org/10.1080/03098770701625761

Danner, D., Hagemann, D., & Funke, J. (2017). Measuring individual differences in implicit learning with artificial grammar learning tasks. *Zeitschrift für Psychologie*.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21. https://doi.org/10.1016/j.intell.2006.02.001

Der, G., & Deary, I. J. (2017). The relationship between intelligence and reaction time varies with age: Results from three representative narrow-age age cohorts at 30, 50 and 69 years. *Intelligence*, *64*, 89–97. https://doi.org/10.1016/j.intell.2017.08.001

Dillon, R. F., Pohlmann, J. T., & Lohman, D. F. (1981). A factor analysis of Raven's Advanced Progressive Matrices freed of difficulty factors. *Educational and Psychological Measurement*, *41*(4), 1295–1302. https://doi.org/10.1177/001316448104100438

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological bulletin*, *145*(5), 508. https://doi.org/10.1037/bul0000192

Draheim, C., Tshukara, J. S., & Engle, R. W. (2024). Replication and extension of the toolbox approach to measuring attention control. *Behavior Research Methods*, *56*(3), 2135–2157. https://doi.org/10.3758/s13428-023-02140-2

Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). *The Wechsler Intelligence Scales and Gf-Gc theory: A contemporary approach to interpretation.* Allyn & Bacon.

Formann, A. K., Piswanger, K., & Waldherr, K. (2011). Wiener Matrizen-Test 2: Ein Rasch-skalierter sprachfreier Kurztest zu Erfassung der Intelligenz. Hogrefe.

Glady, Y., French, R. M., & Thibaut, J. P. (2017). Children's failure in analogical reasoning tasks: A problem of focus of attention and information integration? *Frontiers in Psychology*, *8*, 707.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature neuroscience*, *6*(3), 316-322. https://doi.org/10.1038/nn1014

Gonthier, C., Braver, T. S., & Bugg, J. M. (2016). Dissociating proactive and reactive control in the Stroop task. *Memory & Cognition*, *44*(5), 778–788. https://doi.org/10.3758/s13421-016-0591-1

Gonthier, C., Macnamara, B. N., Chow, M., Conway, A. R. A., & Braver, T. S. (2016). Inducing Proactive Control Shifts in the AX-CPT. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01822

Gonthier, C., & Roulin, J.-L. (2020). Intraindividual strategy shifts in Raven's matrices, and their dependence on working memory capacity and need for cognition. *Journal of Experimental Psychology: General*, *149*(3), 564–579. https://doi.org/10.1037/xge0000660

Gonthier, C., & Thomassin, N. (2015). Strategy use fully mediates the relationship between working memory capacity and performance on Raven's matrices. *Journal of Experimental Psychology: General*, *144*(5), 916–924. https://doi.org/10.1037/xge0000101

Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. Intelligence, 8(3), 179–203. https://doi.org/10.1016/0160-2896(84)90008-4

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

Horn, W. (1983). Leistungsprüfsystem: LPS [Performance Test System]. Göttingen: Hogrefe.

Horn, J. L. (1965) Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities. [Unpublished dissertation, Illinois University]. ProQuest Dissertation Publishing.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253–270. https://doi.org/10.1037/h0023816

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. *Woodcock-Johnson technical manual*, 197-232.

Jarosz, A. F., Raden, M. J., & Wiley, J. (2019). Working memory capacity and strategy use on the RAPM. Intelligence, 77, 101387. https://doi.org/10.1016/j.intell.2019.101387

Jastrzębski, J., Ciechanowska, I., & Chuderski, A. (2018). The strong link between fluid intelligence and working memory cannot be explained away by strategy use. *Intelligence*, *66*, 44–53. https://doi.org/10.1016/j.intell.2017.11.002

Kan, K.-J., Kievit, R. A., Dolan, C., & der Maas, H. van. (2011). On the interpretation of the CHC factor Gc. *Intelligence*, *39*(5), 292–302. https://doi.org/10.1016/j.intell.2011.05.003

Kalra, P. B., Gabrieli, J. D., & Finn, A. S. (2019). Evidence of stable individual differences in implicit learning. *Cognition*, *190*, 199–211. https://doi.org/10.1016/j.cognition.2019.05.007

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*(3), 151–177. https://doi.org/10.1080/1047840X.2016.1153946

Krampen, D., Gold, A., & Schweizer, K. (2020). Does impulsivity Contribute to the Item-Position Effect? *Psychological Test and Assessment Modeling*, *62*(3), 375-385.

Kriegbaum, K., Becker, N., & Spinath, B. (2018). The relative importance of intelligence and motivation as predictors of school achievement: A meta-analysis. *Educational Research Review*, *25*, 120-148. https://doi.org/10.1016/j.edurev.2018.10.001

Krys, K., de Almeida, I., Wasiel, A., & Vignoles, V. L. (2024). WEIRD–Confucian comparisons: Ongoing cultural biases in psychology's evidence base and some recommendations for improving global representation. *American Psychologist.* Advance online publication. https://doi.org/10.1037/amp0001298

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Kyllonen, P., Hartman, R., Sprenger, A., Weeks, J., Bertling, M., McGrew, K., ... & Stankov, L. (2019). General fluid/inductive reasoning battery for a high-ability population. *Behavior Research Methods*, *51*, 507–522. https://doi.org/10.3758/s13428-018-1098-4

Kyllonen, P., Kell, H. (2017). What Is Fluid Intelligence? Can It Be Improved? In: Rosén, M., Yang Hansen, K., Wolff, U. (eds) Cognitive Abilities and Educational Outcomes. Methodology of Educational Measurement and Assessment. Springer, Cham. https://doi.org/10.1007/978-3-319-43473-5_2

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research*, *52*(5), 661–670.

Laidra, K., Pullmann, H., & Allik, J. (2007). Personality and intelligence as predictors of academic achievement: A cross-sectional study from elementary to secondary school. *Personality and individual differences*, *42*(3), 441-451. https://doi.org/10.1016/j.paid.2006.08.001

Laurence, P. G., Mecca, T. P., Serpa, A., Martin, R., & Macedo, E. C. (2018). Eye Movements and Cognitive Strategy in a Fluid Intelligence Test: Item Type Analysis. *Frontiers in Psychology*, *9*, 380. https://doi.org/10.3389/fpsyg.2018.00380

Li, C., Ren, X., Schweizer, K., & Wang, T. (2022). Strategy use moderates the relation between working memory capacity and fluid intelligence: A combined approach. *Intelligence*, *91*, 101627. https://doi.org/10.1016/j.intell.2022.101627

Liu, Y., Zhan, P., Fu, Y., Chen, Q., Man, K., & Luo, Y. (2023). Using a multi-strategy eye-tracking psychometric model to measure intelligence and identify cognitive strategy in Raven's advanced progressive matrices. *Intelligence*, *100*, 101782. https://doi.org/10.1016/j.intell.2023.101782

Lozano, J. H. (2015). Are impulsivity and intelligence truly related constructs? Evidence based on the fixed-links model. *Personality and Individual Differences*, *85*, 192–198. https://doi.org/10.1016/j.paid.2015.04.049

Lu, D., Zhang, H., Kang, C., & Guo, T. (2016). ERPs evidence for the relationship between fluid intelligence and cognitive control. *NeuroReport*, *27*(6), 379–383. https://doi.org/10.1097/WNR.0000000000000547

Lüthold, P., Lao, J., He, L., Zhou, X., & Caldara, R. (2018). Waldo reveals cultural differences in return fixations. *Visual Cognition*, *26*(10), 817–830. https://doi.org/10.1080/13506285.2018.1561567

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence 37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

McGrew, K. S. (2023). Carroll's three-stratum (3S) cognitive ability theory at 30 years: Impact, 3S-CHC theory clarification, structural replication, and cognitive–achievement psychometric network analysis extension. *Journal of Intelligence*, *11*(2), 32 https://doi.org/10.3390/jintelligence11020032

Nagy, G., Ulitzsch, E., & Lindner, M. A. (2023). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. Journal of Computer Assisted Learning, 39(3), 751–766. https://doi.org/10.1111/jcal.12719

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of experimental child psychology*, *162*, 31-38.

Ortiz, S. O. (2015). CHC theory of intelligence. *Handbook of intelligence: Evolutionary theory, historical perspective, and current concepts*, 209-227.

Pahud, O., Rammsayer, T. H., & Troche, S. J. (2018). Elucidating the functional relationship between speed of information processing and speed-, capacity-, and memory-related aspects of psychometric intelligence. *Advances in cognitive psychology*, *13*(1), 3. https://doi.org/10.5709%2Facp-0233-4

Paxton, J. L., Barch, D. M., Racine, C. A., & Braver, T. S. (2008). Cognitive control, goal maintenance, and prefrontal function in healthy aging. *Cerebral cortex*, *18*(5), 1010–1028. https://doi.org/10.1093/cercor/bhm135

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raden, M. J., & Jarosz, A. F. (2022). Strategy Transfer on Fluid Reasoning Tasks. *Intelligence*, *91*, 101618. https://doi.org/10.1016/j.intell.2021.101618

Raven, J. (2000). The Raven's progressive matrices: change and stability over culture and time. *Cognitive psychology*, *41*(1), 1–48.

Raven, J. C., & Raven, J. (2003). Raven Progressive Matrices. In R. S. McCallum (Ed.), *Handbook of Nonverbal Assessment* (pp. 223–237). Springer. https://doi.org/10.1007/978-1-4615-0153-4_11

Raven, J. C., Raven, J., & Court, J. H. (1998). *Advanced Progressive Matrices [Measurement Instrument].* https://www.testzentrale.ch/shop/advanced-progressive-matrices.html

Redick, T. S. (2014). Cognitive control in context: Working memory capacity and proactive control. *Acta Psychologica*, *145*, 1–9. https://doi.org/10.1016/j.actpsy.2013.10.010

Ren, X., Gong, Q., Chu, P., & Wang, T. (2017). Impulsivity is not related to the ability and position components of intelligence: A comment on Lozano (2015). *Personality and Individual Differences*, *104*, 533–537. https://doi.org/10.1016/j.paid.2016.09.007

Ren, X., Schweizer, K., Wang, T., Chu, P., & Gong, Q. (2017). On the relationship between executive functions of working memory and components derived from fluid intelligence measures. *Acta Psychologica*, *180*, 79–87. https://doi.org/10.1016/j.actpsy.2017.09.002

Ren, X., Schweizer, K., Wang, T., & Xu, F. (2015). The prediction of students' academic performance with fluid intelligence in giving special consideration to the contribution of learning. *Advances in Cognitive Psychology*, *11*(3), 97–105. https://doi.org/10.5709/acp-0175-z

Ren, X., Wang, T., Altmeyer, M., & Schweizer, K. (2014). A learning-based account of fluid intelligence from the perspective of the position effect. *Learning and Individual Differences*, *31*, 30–35. https://doi.org/10.1016/j.lindif.2014.01.002

Richmond, L. L., Redick, T. S., & Braver, T. S. (2015). Remembering to prepare: The benefits (and costs) of high working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1764.

Rosenberg, J. M., Beymer, P. N., Anderson, D. J., Van Lissa, C. J., & Schmidt, J. A. (2019). tidyLPA: An R package to easily carry out latent profile analysis (LPA) using open-source or commercial software. *Journal of Open Source Software*, *3*(30), 978.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1–36.

Roth, M., & Herzberg, P. Y. (2008). Psychodiagnostik in der Praxis: State of the Art? Klinische Diagnostik und Evaluation, 1, 5–18.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, *124*(2), 262.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99–144). The Guilford Press.

Schweizer, K. (2006). The fixed-links model for investigating the effects of general and specific processes on intelligence. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 2*(4), 149–160. https://doi.org/10.1027/1614-2241.2.4.149

Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two-dimensional model of APM. *Psychology Science Quarterly*, *51*(1), 47.

Schweizer, K. (2013). A threshold-free approach to the study of the structure of binary data. *International Journal of Statistics and Probability*, *2*(2), 67. https//doi.org/10.5539/ijsp.v2n2p67

Schweizer, K., Reiss, S., Schreiner, M., & Altmeyer, M. (2012). Validity improvement in two reasoning measures following the elimination of the position effect. *Journal of Individual Differences*. *33*(1), 54–61. https://doi.org/10.1027/1614-0001/a000062

Schweizer, K., Ren, X., Wang, T. (2015). A Comparison of Confirmatory Factor Analysis of Binary Data on the Basis of Tetrachoric Correlations and of Probability-Based Covariances: A Simulation Study. In: Millsap, R., Bolt, D., van der Ark, L., Wang, WC. (eds) Quantitative Psychology Research. Springer Proceedings in Mathematics & Statistics, vol 89. Springer, Cham. https://doi.org/10.1007/978-3-319-07503-7_17

Schweizer, K., & Troche, S. (2018). Is the factor observed in investigations on the item-position effect actually the difficulty factor? *Educational and Psychological Measurement*, *78*(1), 46–69. https://doi.org/10.1177/0013164416670711

Schweizer, K., & Troche, S. (2019). The EV Scaling Method for Variances of Latent Variables. *Methodology*, *15*(4), 175–184. https://doi.org/10.1027/1614-2241/a000179

Schweizer, K., Troche, S. J., & Rammsayer, T. H. (2011). On the special relationship between fluid and general intelligence: New evidence obtained by considering the position effect. *Personality and Individual Differences*, *50*(8), 1249–1254. https://doi.org/10.1016/j.paid.2011.02.019

Schweizer, K., Troche, S., Rammsayer, T., & Zeller, F. (2021). Inductive reasoning and its underlying structure: Support for difficulty and item position effects. *Advances in Cognitive Psychology*, *17*(4), 274–283. https://doi.org/10.5709/acp-0336-5

Schweizer, K., Zeller, F., & Reiß, S. (2020). Higher-order processing and change-to-automaticity as explanations of the item-position effect in reasoning tests. *Acta Psychologica*, *203*, 102991. https://doi.org/10.1016/j.actpsy.2019.102991

Snow, R. E. (1978). Eye Fixation and Strategy Analyses of Individual Differences in Cognitive Aptitudes. In A. M. Lesgold, J. W. Pellegrino, S. D. Fokkema, & R. Glaser (Eds.), *Cognitive Psychology and Instruction* (pp. 299–308). Springer US. https://doi.org/10.1007/978-1-4684-2535-2_27

Snow, R. E. (1980). Aptitude Processes. In R. E. Snow, P.-A. Federico, & W. E. Montague (Eds.), *Aptitude, learning, and instruction: Cognitive process analyses of aptitude* (Vol. 1, pp. 27–64). Erlbaum.

Spearman, C. E. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology* 15, 201–292. https://doi.org/10.2307/1412107

Spearman, C. (1927). *The abilities of man.* Macmillan.

SR Research. (2016). Eyelink 1000 plus [Apparatus and software]. https://www.sr-research.com/eyelink1000plus.html

Starr, A., Vendetti, M. S., & Bunge, S. A. (2018). Eye movements provide insight into individual differences in children's analogical reasoning strategies. *Acta Psychologica*, *186*, 18-26.

Sternberg, R.J. Intelligence and culture. In Handbook of Cultural Psychology; Kitayama, S., Cohen, D., Eds.; Guilford Press: New York, NY, USA, 2007; pp. 547–568.

Sternberg, R. J. (2019). A theory of adaptive intelligence and its relation to general intelligence. *Journal of Intelligence*, *7*(4), 23.

Sternberg, R. J. (2021). *Adaptive intelligence. Surviving and Thriving in Times of Uncertainty.* Cambridge University Press.

Sun, S., Schweizer, K., & Ren, X. (2019). Item-position effect in Raven's matrices: A developmental perspective. *Journal of Cognition and Development*, *20*(3), 370–379. https://doi.org/10.1080/15248372.2019.1581205

Thibaut, J. P., & French, R. M. (2016). Analogical reasoning, control and executive functions: A developmental investigation with eye-tracking. *Cognitive Development*, *38*, 10–26. https://doi.org/10.1016/j.cogdev.2015.12.002

Thurstone, L.L. (1947). *Multiple factor analysis.* University of Chicago Press: Chicago.

Troche, S. J., Wagner, F. L., Schweizer, K., & Rammsayer, T. H. (2016). The Structural Validity of the Culture Fair Test Under Consideration of the Item-Position Effect. *European Journal of Psychological Assessment*, *35*(2), 182–189. https://doi.org/10.1027/1015-5759/a000384

Troche, S. J., von Gugelberg, H. M., Pahud, O., & Rammsayer, T. H. (2021). Do executive attentional processes uniquely or commonly explain psychometric g and correlations in the positive manifold? A structural equation modeling and network-analysis approach to investigate the process overlap theory. *Journal of Intelligence, 9*(3), 1–17. https://doi.org/10.3390/jintelligence9030037

van Der Maas, H. L., Kan, K. J., Marsman, M., & Stevenson, C. E. (2017). Network models for cognitive development and intelligence. *Journal of Intelligence*, *5*(2), 16. https://doi.org/10.3390/jintelligence5020016

van der Ven, A., & Ellis, J. L. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences*, *29*(1), 45–64. https://doi.org/10.1016/S0191-8869(99)00177-4 von Gugelberg, H. M., Schweizer, K., & Troche, S. J. (2021). The dual mechanisms of cognitive control and their relation to reasoning and the item-position effect. *Acta Psychologica*, *221*, 103448. https://doi.org/10.1016/j.actpsy.2021.103448

von Gugelberg, H. M. (2022). bindabox: Toolbox for fixed-links modelling. Version 1.1.0 https://github.com/hvongbg/bindabox/

von Gugelberg, H. M. & Troche, S. J. (2022a). Improving Measurement of Reasoning Ability Through Consideration of the Item-Position Effect. *In C. Bermeitinger & W. Greve (Eds.) 52nd Congress of the German Psychological Society* (p. 438). Pabst Science Publishers.

von Gugelberg, H. M. & Troche, S. J. (2022b). Switching and Fluid Intelligence: A Fixed-Links Approach. *Presented at the Doctoral Program Brain and Behavioral Sciences Summer Course.* June 13th – June 16th, 2022

von Gugelberg, H. M., Schweizer, K., & Troche, S. J. (2025). Experimental evidence for rule learning as the underlying source of the item-position effect in reasoning ability measures. *Learning and Individual differences. 118,* 102622. https://doi.org/10.1016/j.lindif.2024.102622

von Gugelberg, H. M. & Troche, S. J. (*in preparation*). Individual Differences in Strategy and the Item-Position Effect in Reasoning Ability Measures.

Verguts, T., & De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement*, *24*(2), 151–162. https://doi.org/10.1177/01466210022031589

Verguts, T., & De Boeck, P. (2002). The induction of solution rules in Raven's Progressive Matrices Test. *European Journal of Cognitive Psychology*, *14*(4), 521–547. https://doi.org/10.1080/09541440143000230

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, *34*(3), 261–272. https://doi.org/10.1016/j.intell.2005.11.003

Vigneau, F., & Bors, D. A. (2008). The quest for item types based on information processing: An analysis of Raven's Advanced Progressive Matrices, with a consideration of gender differences. *Intelligence*, *36*(6), 702–710. https://doi.org/10.1016/j.intell.2008.04.004

Wang, T., Zhang, Q., & Schweizer, K. (2020). Investigating the item-position effect in a longitudinal data with special emphasis on the information provided by the variance parameter. *Psychological Test and Assessment Modeling*, *62*(3), 404-417.

Wechsler, D. (2017). *Wechsler Intelligence Scale for Children – Fifth Edition*. Pearson.

Weiss R. H. (2006). *Grundintelligenztest Skala 2 – Revision.* Hogrefe.

Wilhelm, O., & Kyllonen, P. (2021). To predict the future, consider the past: Revisiting Carroll (1993) as a guide to the future of intelligence research. *Intelligence*, *89*, 101585. https://doi.org/10.1016/j.intell.2021.101585

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Test review. *Rehabilitation counseling bulletin*, *44*(4), 232–235.

Zeller, F., Krampen, D., Reiss, S., & Schweizer, K. (2017). Do adaptive representations of the item-position effect in APM improve model fit? A simulation study. *Educational and Psychological Measurement*, *77*(5), 743–765. https://doi.org/10.1177/00131644166549

Zeller, F., Reiss, S., & Schweizer, K. (2017). Is the item-position effect in achievement measures induced by increasing item difficulty? *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 745–754. https://doi.org/10.1080/10705511.2017.1306706

Neither a lofty degree of intelligence nor imagination nor both together go to the making of genius. Love, love, love, that is the soul of genius.

Wolfgang Amadeus Mozart

APPENDIX

Reasoning ability measures, omnipresent, yet not fully understood. A closer look at the learning hypothesis.

Anhang für die Inauguraldissertation der Philosophisch-humanwissenschaftlichen Fakultät der Universität Bern zur Erlangung der Doktorwürde vorgelegt von:

Helene M. von Gugelberg

Maienfeld, September 2024

Table of Contents

Appendix A

Study 1: Attentional Control and the Learning Hypothesis

This article is published as:

von Gugelberg, H. M., Schweizer, K., & Troche, S. J. (2021). The dual mechanisms of cognitive control and their relation to reasoning and the item-position effect. *Acta Psychologica*, *221*, 103448. https://doi.org/10.1016/j.actpsy.2021.103448

Acta Psychologica 221 (2021) 103448

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

The dual mechanisms of cognitive control and their relation to reasoning and the item-position effect☆

Helene M. von Gugelberg[a,*], Karl Schweizer[b], Stefan J. Troche[a]

[a] Department of Psychology, University of Bern, Switzerland
[b] Department of Psychology and Sports Sciences, Goethe University Frankfurt, Frankfurt a. M, Germany

ARTICLE INFO

Classification codes:
2223
2340
3120

Keywords:
Dual mechanism of cognitive control
Fluid intelligence
Item-position effect
Reasoning
AX-CPT

ABSTRACT

Braver's (2012) *dual mechanisms of cognitive control* differentiate between proactive control (PMC; i.e. early selection and maintenance of goal-relevant information) and reactive control (RMC; i.e. a late mobilization of attention when required). It has been suggested that higher cognitive capacities (as indicated by reasoning ability as a major characteristic of fluid intelligence) facilitate using the more resource-demanding PMC. We propose the following alternative explanation: engagement in PMC during the completion of reasoning tests leads to better test performance because gained knowledge (i.e. rules learned) during completion of early items is better maintained and transferred to later items. This learning of rules during the completion of a reasoning test results in an item-position effect (IPE) as an additional source of individual differences besides reasoning ability. We investigated this idea in a sample of 210 young adults who completed the AX-Continuous Performance Task (AX-CPT) and the Vienna Matrices Test (VMT). Using fixed-links modeling, we separated an IPE from reasoning ability in the VMT. Based on reaction time (RT) patterns across AX-CPT conditions, we identified three different groups by means of latent-profile analysis. RT patterns indicated engagement in PMC for Group A, mixed PMC and RMC for Group B, and RMC for Group C. With the consideration of the IPE, groups did not differ in their reasoning abilities. However, Group A (engaging in PMC) had a more pronounced IPE than Group C (engaging in RMC). Therefore, we conclude that PMC contributes to a stronger IPE, which in turn leads to higher scores in reasoning tests as measures of fluid intelligence.

1. Dual mechanisms of cognitive control

The ability to control our behaviour in order to achieve our goals is an important ability to master everyday life. We can plan into the future and suppress actions when anticipating future consequences (e.g. Sodian & Frith, 2008). The ability to register and maintain context information is assumed to play a crucial role and is also often referred to as cognitive control or attention control (Paxton et al., 2008).

Within the framework of the dual mechanisms of cognitive control (DMC), Braver (2012) put forward the idea that context representation and maintenance during information processing are the key components of cognitive control. As the name implies, there are two distinguishable mechanisms of cognitive control (Braver, 2012). Maintaining goal-relevant information in anticipation of a certain event or stimulus is referred to as a *Proactive Mechanism of Control* (PMC). This mechanism of control means early selection and maintenance of goal relevant information in anticipation of a challenging event in order to ideally guide attention. A *Reactive Mechanism of Control* (RMC), on the other hand, describes stimulus or event driven activation of goal-relevant information. With this mechanism of control, specific information is processed when it appears, but not anticipated to prepare processing in advance. This can be seen as late correction of past occurrences, as this mechanism depends on the occurrence of a specific event rather than its anticipation. It was suggested that RMC places less demands on cognitive resources than PMC, which is rather cognitively demanding (Braver, 2012).

Evidence in favour of two dissociable mechanisms of cognitive control stems from different areas of research. For example, neurophysiological studies provided evidence for different brain areas associated with PMC and RMC (Braver et al., 2009; Paxton et al., 2008). On

☆ Initial results of the latent profile analysis and the multilevel model analysis were presented at the 63rd conference for experimental psychologists (2021, TeaP, Ulm).

* Corresponding author.
E-mail address: helene.vongugelberg@psy.unibe.ch (H.M. von Gugelberg).

https://doi.org/10.1016/j.actpsy.2021.103448
Received 8 September 2021; Received in revised form 5 November 2021; Accepted 9 November 2021
Available online 13 November 2021

the behavioural level, Gonthier, Braver, and Bugg (2016) analysed reaction times (RTs) of different variations of the Stroop task and showed that the effects of the two mechanisms of control could be dissociated by experimental manipulation. Furthermore, Braver et al. (2001) reported that young adults showed more PMC than RMC while the opposite was found in older adults. It should be noted, however, that – although using PMC seems to be more advantageous than RMC – successful cognition is assumed to depend on a mixture of both mechanisms (Braver, 2012).

2. Dual mechanisms of cognitive control and intelligence

Burgess and Braver (2010) observed RMC-related brain activity when interference expectancy was low but an increase in PMC-related brain activity when interference expectancy was high in a recent-probes task. These results indicated that individuals shifted from one mechanism of cognitive control to the other when the situation required such a shift and cognitive capacities were available. In their behavioural data, Burgess and Braver (2010) also compared individuals with high and low fluid intelligence with fluid intelligence (Gf) defined as the ability to solve novel problems (Jensen, 1998). Overall high Gf individuals outperformed low Gf individuals in the recent-probes task (Burgess & Braver, 2010).

Gray et al.'s (2003) investigation of neural mechanisms of Gf lead to a similar observation. These authors applied an n-back task and systematically varied the amount of interference between conditions. Results showed stronger event-related neural activity in brain areas associated with PMC in the high interference condition. Most importantly for the present purpose, individuals with higher Gf did not only outperform individuals with lower Gf in the high interference condition but also showed stronger PMC-related brain activity. These results suggested that high Gf individuals engaged more strongly in PMC than low Gf individuals, which might be the reason for their better performance when being confronted with high interference (Gray et al., 2003).

The Gf-related differences reported by Burgess and Braver (2010) were not larger in the high than in the low interference condition, which was the case in the study of Gray et al. (2003). However, Burgess and Braver (2010) reported that, in a pilot study, high Gf individuals were indeed less affected by interference than low Gf individuals, especially when the inference expectancy was high.

These previous results on the relationship between Gf and the dual mechanisms of cognitive control (Burgess & Braver, 2010; Gray et al., 2003) have been taken as evidence that the higher cognitive capacities of individuals with high Gf facilitate or enable the use of PMC. Individuals with lower Gf, on the other hand, are more likely to engage in the less capacity-demanding RMC (Braver, 2012). The aim of the present study was to investigate an alternative explanation of the link between Gf and the dual mechanisms of cognitive control, which assumes a reversed direction of the effect. More specifically, we assumed that the engagement in PMC in contrast to RMC during the completion of a reasoning test leads to better performance on a reasoning test (and thereby to a higher estimation of Gf). To substantiate this assumption, we will outline in the following paragraphs how performance on reasoning tests is not only influenced by reasoning ability but also by an item-position effect (IPE) and how this IPE might be influenced by the use of PMC/RMC.

3. The item-position effect

Both previous studies on Gf and the dual mechanisms of cognitive control (Burgess & Braver, 2010; Gray et al., 2003) assessed Gf with Raven's Advanced Progressive Matrices (APM; Raven & Raven, 2003). With this type of test, participants have to identify a rule within a presented matrix per item and use this rule to choose one out of eight alternatives to fill the empty cell in the matrix correctly. Such psychometric reasoning tests are well-established and valid measures of Gf (Gustafsson, 1984; Kan et al., 2011; Schweizer et al., 2011) since reasoning ability is the main component of Gf (Carroll, 1993). However, there is also growing evidence that these reasoning tests are not homogeneous and therefore are no pure measures of Gf or reasoning ability. Confirmatory factor analyses (CFA) on the items of reasoning tests and primarily on the APM (e.g. Sun et al., 2019; Zeller et al., 2017) pointed to an IPE. This IPE could be dissociated from reasoning ability by means of bifactor measurement models, in which the factor loadings of the latent variable representing the IPE were fixed to increase monotonically from the first to the last item. The IPE explained a substantial portion of individual differences in test performance in addition to the latent variable reflecting reasoning ability and also improved the measurement model substantially (Schweizer, 2013; Troche et al., 2016). The IPE indicates that the processing of earlier items influences the processing of later items and the strength of this influence varies strongly between individuals, which is depicted by the amount of variance of the latent variable representing the IPE. At first sight, it might be assumed that the IPE just reflects the increasing item difficulty in a reasoning scale. This explanation could be ruled out with simulation studies (Schweizer & Troche, 2018) and empirical studies (Zeller et al., 2017). For example, in the study by Zeller et al. (2017) items were presented in random order. This manipulation of item order led to a dissociation of item position and item difficulty, and the IPE could still be clearly observed but not anymore explained by item difficulty. Importantly, the two components of the APM (i.e. reasoning ability and IPE) were not only separable on a statistical level, but also showed different correlations with several psychological constructs. Ren et al. (2017), for example, highlighted that when IPE and reasoning ability were both being considered, reasoning ability was moderately related to updating and inhibition, while the IPE was associated with updating and shifting abilities but not with inhibition.

To date, the most plausible explanation for the IPE is, that it reflects the learning of rules underlying the matrices during the processing of an item series (Ren et al., 2014; Ren et al., 2015; Sun et al., 2019; Zeller et al., 2017). This explanation is based on the finding that the IPE but not reasoning ability was strongly related to complex learning (Ren et al., 2014). According to the learning hypothesis, the underlying rules have to be identified and correctly applied to correctly solve a reasoning test item. If an individual can successfully carry over this newly gained knowledge to the next items, solving the next items can benefit increasingly from the processing of previous items. It is reasonable that individuals do not only differ in their ability to detect the rules underlying the matrices but also in their ability to use knowledge gained during the solving of earlier items or, stated differently, in their ability to use context information when an item is seen as an element of an item series. It is this ability, which is assumed to underlie individual differences in the IPE.

4. Item-position effect and cognitive control

The insights into the meaning of the IPE also provide a functional link with the dual mechanisms of cognitive control since the core of PMC is the use of context information to ideally guide attention during current information processing. Individuals engaging in PMC would already have previously learned rules on hand. Their first inspection of an item would already include the direct comparison of a new item with the experience from previous items. Since they show a disposition that supports maintenance of information, they are less likely to miss a connection or loose trace of a rule already learned. This should lead to a clear and increasing advantage when solving a series of similar items.

On the contrary, an individual engaging primarily in RMC might be expected to first process each reasoning item separately without taking previous experience into consideration, and only then accesses prior experience during previously solved items. These individuals would only benefit from prior knowledge, if a rule is correctly detected during the first inspection and the connection between this rule and an earlier rule can be made. This approach is less likely to be successful, since it depends on the successful retrieval and connection of previously applied rules.

In other words, individuals using PMC might benefit from this mechanism of cognitive control during experimental tasks as well as during the completion of a reasoning test, while individuals using RMC do not. This would lead to a positive correlation between performance on the experimental task assessing PMC/RMC and the reasoning task that is not due to reasoning ability but the IPE in the reasoning task. This would suggest, the previously observed relation between higher Gf as measured by a reasoning test and engagement in PMC by Braver and his colleagues (Burgess & Braver, 2010; Gray et al., 2003) can be interpreted in different ways. One interpretation refers to the original explanation that high compared to low Gf individuals possess higher cognitive capacities, which facilitate engagement in PMC during cognitively challenging situations (Burgess & Braver, 2010; Gray et al., 2003). Alternatively, it might be possible that individuals differ in their extent of engagement in PMC and that stronger engagement in PMC has a positive influence on the learning of rules and their later application when completing a reasoning test. This should become evident in a stronger IPE (rather than higher reasoning ability) resulting in better performance on the reasoning test.

5. Current research

The goal of the present study was to investigate this alternative explanation. More specifically, the goal was to examine, whether individuals that have a predisposition to engage in PMC differ from individuals that have a predisposition to engage in RMC during their performance on a reasoning test, due to higher reasoning abilities (as an indicator of Gf and, thus, of cognitive capacities) or due to a more pronounced IPE.

For this purpose and to identify individuals using PMC and or RMC, we used the AX-Continuous Performance Task (AX-CPT) paradigm (e.g., Gonthier, Macnamara, et al., 2016). In each trial of the AX-CPT, participants are presented with a cue letter followed by a probe letter (see Fig. 1). The task has four conditions, which differ in the combination of cue and probe letters. If the cue letter “A” is followed by the probe letter “X” (AX condition), participants are supposed to give a target response, by pressing a designated button with the right index finger. For all other cue-probe combinations, a non-target response is required, and participants are instructed to press another button with the left index finger. These conditions are often abbreviated as BX condition, BY condition and AY condition. Whereas “B” always indicates any letter but “A” as cue, “Y” any letter but “X” as probe, and the letters “X” and “A” represent themselves as probe or cue respectively.

Several studies mentioned RT differences between single conditions of the AX-CPT (e.g. Braver et al., 2001; Gonthier, Macnamara, et al., 2016; Paxton et al., 2008; Redick, 2014) which were interpreted as markers or identifiers of a certain mechanism of control. Overall, individuals strongly engaging in PMC should give a nearly immediate response upon appearance of the probe in the BX and BY condition as the cue letter holds sufficient information to prepare a correct non-target response. Also, when only applying PMC no significant RT difference between the BX and BY condition should arise. The target response for the AX condition should be somewhat slower, as the individual has to wait for the probe to appear, since it is relevant for the response. RTs in the AY condition should also be notably slower when compared to the BX and BY condition, since the appearance of the probe letter has to be awaited before giving a correct response.

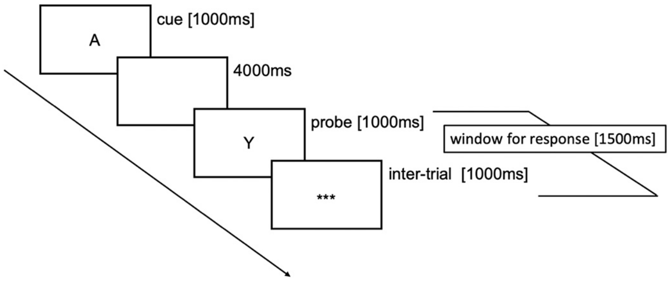


Fig. 1. Simplified display of one trial of the AY condition of the AX-CPT.

For individuals applying predominantly RMC, a different RT pattern should emerge, since responses are only formed after the probe has been presented. These individuals would give their fastest response in the BY and AY conditions, since the probe letter Y contains all information necessary to respond and no further processing of the cue letter is necessary. Additionally, there is no reason for a difference in RT between these two conditions. In the AX and BX conditions, RTs should be longer because the cue letter has to be retrieved after the probe letter X has been presented. Only then a response can be prepared. Since AX trials are presented more frequently, a target response could have a small advantage when compared to a non-target response. This advantage would be noticeable in faster RTs in the AX condition when compared to RTs in the BX condition.

For the present study, we expected that, in line with Burgess and Braver (2010), individuals showing an RT pattern with all the markers described above for PMC would achieve higher test scores on a reasoning test than individuals with an RT pattern coinciding with the markers described for RMC. However, when dissociating the IPE from reasoning ability, we expected that this PMC-related advantage would be obvious in a more pronounced IPE rather than in higher reasoning ability. There were several obstacles to investigating this idea. We had to first identify individuals who show a disposition to engage in PMC or RMC according to their RT pattern in the AX-CPT. Additionally, correlational analyses between RTs in specific conditions of the AX-CPT and reasoning test scores would be difficult to interpret since shorter RTs are consistently related to higher reasoning test scores regardless of the specific processes for which RTs are obtained (e.g. Der & Deary, 2017). Engaging in PMC or RMC, however, should lead to different variations of RTs across the four AX-CPT conditions (i.e. different RT patterns) and not only in faster RTs per se. Therefore, to identify possible underling groups of individuals that show similar dispositions in their use of cognitive control when completing the AX-CPT, we applied latent profile analyses (LPA). This approach enabled us to detect different groups without coercing certain structures (e.g., assuming exactly two groups) based on theoretical assumptions. Grouping individuals by means of LPA ensured that the groups were allowed to vary in their RT pattern. The LPA proved to be an objective approach to identifying unique groups of individuals showing different RTs during the completion of the AX-CPT, which also facilitates the replication in future studies. To characterize the identified groups in terms of PMC/RMC, multilevel modeling (MLM) was applied to analyse RT differences within the groups (between the AX-CPT conditions) and compare these RT differences between groups (i.e. cross-level interactions). In a last step, we investigated whether an IPE could be extracted in addition to reasoning ability from a reasoning test and whether the groups differed in their factor scores on reasoning ability and/or IPE. With this procedure the overarching objective could be specified by the following research questions:

1. Do the groups identified by means of LPA show RT patterns across the four conditions of AX-CPT that coincide with the markers assumed for PMC or RMC?
2. Can the IPE be detected in the reasoning test scores of the present sample in addition to a latent variable representing reasoning ability?
3. Do individuals that show the most PMC-consistent RT patterns have higher reasoning abilities and/or a more pronounced IPE than individuals with RMC-consistent RT patterns?

6. Method

6.1. *Participants*

A total of 210 individuals participated in the present study. While 161 participants described themselves as female, and 48 as male, one participant did not declare gender, age nor highest level of education. Mean age of the sample was 22.4 years (SD = 4.6 years). One hundred sixty-nine participants reported having university entrance qualification as their highest level of education, 21 a Bachelor's degree or higher, and 19 participants had neither. All participants reported normal or corrected-to-normal vision and gave written informed consent. The study protocol was approved by the local ethic committee of the University of Witten/Herdecke (No. 175/2017).

6.2. *Vienna matrices test*

The Vienna Matrices Test (VMT; Formann et al., 2011) is a measure of Gf, similar to Raven's APM, and consists of 18 items. Each item contains a 3 × 3 matrix, with each cell containing a geometric figure. The right cell on the bottom right is filled with a question mark. Participants are instructed to choose one out of eight possible response alternatives, which completes the matrix according to the underlying rule when substituting the question mark. In line with the manual, no time limitation was used and each participant gave a response to each item.

According to the manual, Cronbach's Alpha is approximately $\alpha = 0.80$. Each item was coded with 1 or 0 when a correct or an incorrect response was given, respectively. To obtain information on the representativeness of the sample regarding fluid intelligence, correct responses were summed up and transformed into age-stratified IQ scores as suggested by the manual.

6.3. *AX continuous performance task*

6.3.1. *Apparatus and stimuli*

The AX-CPT was adapted from Gonthier, Macnamara, et al. (2016) and programmed with E-Prime 2.0 Software. Participants completed the task on a Lenovo Thinkpad T510 with a 15.5″ monitor, which was positioned approximately 50 cm from participants' eyes. Responses were given via an external Cedrus response pad (Model RB-830; Cedrus Coporation; n.d.) with a registration accuracy of ±1 ms. Stimuli were black letters presented in the centre of the white monitor. Each letter had a height of 1 cm and a width of 0.8 cm.

6.3.2. *Procedure*

The task consisted of four conditions (AX-, AY-, BX-, and BY condition). In the 80 trials of the AX condition, the cue letter A was followed by the probe letter X. The AY condition contained 20 trials with the letter A as cue and any letter but X as probe. In the 80 trials of the BY condition, cue and probe letters were neither A nor X. In the 20 trials of the BX condition, the cue was any letter but A and the probe letter was X. The trials of the four conditions were presented in random order.

Each trial started with the cue letter presented for 1000 ms, followed by a blanc screen lasting 4000 ms and then the probe letter was presented for 1000 ms (see Fig. 1). Afterwards, three black asterisks were presented in the centre of the screen for 1000 ms before the next trial started. Participants were instructed to press a designated key with the right forefinger in response to trials from the AX condition and to press another designated key with the left forefinger in response to trials from the three other conditions. The instructions emphasized speed but to avoid errors. As dependent variable, mean RT of correct responses given within 150 to 1500 ms after the onset of the probe was recorded for each of the four conditions.

The task was preceded by written instructions and 10 practice trials to ensure that participants had understood the instructions. The duration of the task was approximately 20 min.

6.4. *Time course of the study*

In a first session, the VMT was completed as paper-pencil test in groups of two to five participants. In this session, further tests were administered, which are irrelevant for the present purpose. The second (individual) session took place within four to seven days after the first session, where each participant completed the AX-CPT followed by two other experimental tasks.

6.5. *Statistical analysis*

All analyses were run with R software using the packages *tidyLPA* (Rosenberg et al., 2019), *lmerTest* (Kuznetsova et al., 2017), *lavaan* (Rosseel, 2012), and *MBESS* (Kelley, 2007).

6.5.1. *Identification of groups*

In order to identify different groups according to the RT patterns across the four conditions of the AX-CPT, the mean RT for each participant in each condition was calculated and submitted to a latent profile analysis that used an expectation–maximization algorithm. Four types of models were computed. The four LPAs differed from each other by the assumption of equal (Model 1 and 3) or varying variances (Model 2 and 4) and by the assumption of zero covariances (Model 1 and 2) or varying covariances (Model 3 and 4). For each model, solutions for two up to eight possible groups were calculated resulting in 32 solutions. The best solution was identified by an analytic hierarchy process (AHP, Akogul & Erisoglu, 2017). The AHP took the information of various fit indices (AIC, AWE, BIC, CLC, KIC, see Table 1) into account and inverted their values to create a decision matrix, whereof it computed a composite relative importance vector (C-RIV) for each solution. According to Akogul and Erisoglu (2017), the solution with the highest C-RIV should be regarded as the best solution.

6.5.2. *Group characteristics*

In order to examine whether the response patterns of the identified groups could be distinguished, we applied multilevel modeling (MLM) to analyse RT differences between and within the identified groups for all four conditions of the AX-CPT. As our hypothesis would be reflected in cross-level interactions (different slopes between groups, meaning different RT differences between groups and conditions) a Slope-as-Outcome model[1] with group affiliation (Level 2) and condition (Level

Table 1
Mean and standard deviation (in parentheses) of IQ and reaction times (RT in milliseconds) in the four AX-CPT conditions for the full sample as well as the subsamples identified by the latent profile analysis.

	VMT raw scores	IQ scores[a]	RT_{AX}	RT_{AY}	RT_{BX}	RT_{BY}
Full sample	13.69	98.40	408	507	393	384
(N = 210)	(3.06)	(14.34)	(99)	(100)	(139)	(129)
Group A (n	14.19	100.75	357	445	305	307
= 114)	(2.92)	(13.62)	(30)	(40)	(34)	(31)
Group B (n	13.61	97.98	416	532	418	400
= 67)	(2.93)	(13.70)	(48)	(55)	(56)	(53)
Group C (n	11.90	90.07	594	693	683	654
= 29)	(3.30)	(15.77)	(130)	(88)	(110)	(115)

[a] IQ calculations were based on age-based norms, therefore the information of one participant in the full sample as well as Group B is missing.

[1] Complete equation of the slope-as-outcome model Aa in Table 3: $RT_{ij} = \gamma_{00} + \gamma_{01}GroupB + \gamma_{02}GroupC + \gamma_{10}AY + \gamma_{20}BX + \gamma_{30}BY + \gamma_{11}AY{:}GroupB + \gamma_{21}BX{:}GroupB + \gamma_{31}BY{:}GroupB + \gamma_{12}AY{:}GroupC + \gamma_{22}BX{:}GroupC + \gamma_{32}BY{:}GroupC + \varepsilon_{ij} + \upsilon_{0j} + \upsilon_{1j}$. With *i* indicating the individual within a Group and *j* the Group, υ_{0j} the random effects of the intercept, υ_{1j} the random effects of the slope, ε_{ij} the residual variance.

1) as a predictor of RT was calculated. Here, cross-level interactions can be seen as an indicator of different engagement in cognitive control. Models were calculated with Restricted Maximum Likelihood estimation (REML). This is preferable, as it is less prone to Type I errors compared to Maximum Likelihood estimation and well suited for small groups ($n < 50$; McNeish, 2017).

6.5.3. Identification of an item-position effect

For the separation of the IPE from reasoning ability, the 18 items of the VMT were analysed by a CFA using the robust maximum likelihood estimation. In a first (congeneric) model, one latent variable was derived from the 18 items with all factor loadings being freely estimated. This latent variable is assumed to reflect reasoning ability as an indicator of Gf. In a next step, the IPE was added to this model as a second latent variable. The correlation between the two latent variables (reasoning ability and item-position effect) was set to zero in order to avoid overlap of the variances. The factor loadings of the second latent variable were fixed to describe a quadratic increase from the first to the last item according to the following equation (cf. Troche et al., 2016):

$$f(i) = \frac{i^2}{k^2}$$

In this equation *i* represents the position of a given item, *k* the total number of items in the test, and *f* (*i*) the factor loading calculated for item *i*. This enables to account for the increasing variance appearing within the items throughout test completion. The gap between the distribution of binary manifest data and normal distribution of the latent variables was bridged by weighting each factor loading with the standard deviation of the respective item (Schweizer, 2013). The statistical significance of the variance of the latent variable representing the IPE was tested to investigate whether the IPE indeed represented a substantial amount of variance in the VMT items. The congeneric and the bifactor model were evaluated by means of model fit indices (χ^2, SRMR, RMSEA, CFI). As recommended by DiStefano (2016), values below 0.06 and 0.08 for the Root Mean Squared Error of Approximation (RMSEA) and for the Standardized Root Mean Square Residual (SRMR), respectively, indicated a good model/data fit. Further, a χ^2/df ratio of less than 2 (Wang & Wang, 2020) and a Comparative Fit Index (CFI) larger than 0.95 were regarded as evidence for a good fit. Models were compared by means of the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) where lower values indicate better fit.

6.5.4. Relation of item position effect, reasoning and cognitive control

In a final step, the groups identified by the LPA were compared regarding differences in reasoning ability and the IPE. For this purpose, factor scores for the reasoning ability and the IPE were extracted. Factor scores depict for each individual the standing on the latent variable in relation to the whole sample. Factor scores are *z* standardized, therefore interpretation of values are always in relation to the mean of the whole sample. Afterwards, the factor scores were compared between the groups by means of pairwise independent *t*-tests. Data used for this analysis can be requested from the corresponding author.

7. Results

For the analysis of RTs, observations below 150 ms and above 1000 ms were excluded (1.39% of all observations). As in Gonthier, Macnamara, et al. (2016), only correct answers were included, this reduced the total of observations by another 2.17%. Table 1 gives descriptive statistics of RTs in the four conditions of the AX-CPT for the full sample (first row). Also reported in Table 1 are means and standard deviations of VMT raw scores and IQ scores. The IQ scores were close to the mean of 100 and the standard deviation of 15 in the representative norm sample reported in the manual of the VMT. Cronbach's alpha was $\alpha = 0.75$, which was close to $\alpha = 0.80$ as reported in the manual.

7.1. Identification of groups

To identify whether different groups of individuals can be found within the RT data, LPAs for the four types of models were run. For each model, the fit indices for solutions with two up to eight groups were computed (see Table 2). The above mentioned AHP (Akogul & Erisoglu, 2017) was used to determine the best solution. According to this process, a model with three groups with variances and covariances allowed to vary between groups and conditions yielded the best description of the data.

This solution assigned 114 participants to Group A, 67 participants to Group B, and 29 participants to Group C. The RT patterns across the four AX-CPT conditions of the three groups are given in Table 1 and are illustrated in Fig. 2. This solution also made sense from a theoretical point of view, as the additional groups in solutions with more than three groups showed response patterns, which were similar to and overlapping with the response patterns of the groups identified in the solution with three groups.

To describe the three groups according to their RT patterns across the four AX-CPT conditions (see Fig. 2), a MLM analysis was conducted. The fully unconditional intercept-only model revealed that participant effects explained 71% of the variance in the RTs as indicated by the intraclass correlation coefficient (ICC[2] = 0.714).

7.2. Group characteristics

To compare all conditions between (3 × 3 intercepts, 3 × 6 slopes) and within groups (3 × 6), a total of nine Slope-as-Outcome models had to be calculated, releveling the group or condition variable for each model. Detailed information about the calculated models is presented in Tables 3, 4, and 5. With releveling we were interested in a total of 45 comparisons. To avoid alpha inflation, we used the conservative Bonferroni correction and adjusted alpha to $\alpha = 0.0011$.

In Models Aa, Ab, and Ac (Table 3) Group A and the AX, AY and BX condition represent the intercept, respectively. In Models Ba, Bb and Bc, intercepts were again the AX, AY, and BX conditions, respectively, but for Group B (see Table 4). Finally, the intercept was releveled to Group C and the AX, AY, and BX condition, respectively (Models Ca to Cc in Table 5). Most relevant results are highlighted below while the full information can be taken from the tables.

When comparing the RT differences between conditions, Group A showed the strongest similarity to the RT pattern expected for individuals using PMC. Group A had significantly faster RTs in the BX and BY conditions compared to the AY and, most importantly, to the AX condition. Noteworthy is also, that the difference between RTs in the BX and in the AY condition, which has been interpreted as a strong indicator for PMC by Braver et al. (2001), was significantly larger in Group A than in the other two groups. RTs in the BX and BY conditions did not differ from each other in Group A, which was another marker for PMC.

In Group B, RTs in the AY condition were significantly longer than RTs in the other three conditions. RTs being longer in the AY condition when compared to the BX and the BY condition, fit the predicted outcome for individuals engaging in PMC. Group B showed similar RTs in the AX condition as in the BX and BY conditions. This did neither fit the assumptions made for PMC nor RMC, since with PMC responses in the BX and BY condition should be the fastest, and with RMC the RT of the BY and AX condition should be significantly different. Interestingly, in Group B (and in contrast to Group A), RTs in the BX condition were significantly longer than in the BY condition, which was consistent with the assumptions made for RMC. While the difference between the BX and BY condition was significant, it did not significantly differ from the

[2] Intraclass Correlation Coefficient (ICC) is defined as the Level-2 variance (participant effects) in proportion to the overall variance of the dependent variable (reaction time) in the intercept-only model: $ICC = {\sigma_{\upsilon 0}}^2/({\sigma_{\upsilon 0}}^2 + {\sigma_\varepsilon}^2)$

Table 2
Fit indices for all the estimated models by the latent profile analysis. Model number indicates model type and Groups the number of groups set for the estimation.

Model	Groups	AIC	AWE	BIC	CLC	KIC	C-RIV
1	1	10,371.05	10,462.61	10,397.83	10,357.05	10,382.05	0.02721
1	2	9617.78	9767.83	9661.29	9593.75	9633.78	0.02930
1	3	9276.20	9484.76	9336.44	9242.13	9297.20	0.03034
1	4	9075.63	9342.74	9152.62	9031.50	9101.63	0.03097
1	5	8969.77	9295.38	9063.49	8915.60	9000.77	0.03129
1	6	8978.42	9362.70	9088.88	8914.05	9014.42	0.03122
1	7	8988.38	9431.30	9115.57	8913.85	9029.38	0.03115
1	8	8998.08	9499.61	9142.00	8913.40	9044.08	0.03108
2	1	10,371.05	10,462.61	10,397.83	10,357.05	10,382.05	0.02721
2	2	9424.59	9621.50	9481.49	9392.47	9444.59	0.02987
2	3	9058.05	9360.24	9145.08	9007.92	9087.05	0.03100
2	4	8876.24	9283.64	8993.39	8808.14	8914.24	0.03156
2	5	8776.47	9289.10	8923.74	8690.38	8823.47	0.03185
2	6	8751.00	9368.94	8928.39	8646.85	8807.00	0.03187
2	7	8741.91	9465.18	8949.43	8619.69	8806.91	0.03183
2	8	8742.14	9570.62	8979.78	8601.94	8816.14	0.03176
3	1	8982.36	9144.08	9029.22	8956.36	8999.36	0.03136
3	2	8831.93	9052.14	8895.53	8795.91	8853.93	0.03185
3	3	8799.38	9078.23	8879.71	8753.19	8826.38	0.03193
3	4	8809.78	9147.71	8906.85	8752.99	8841.78	0.03185
3	5	8783.56	9179.60	8897.36	8717.12	8820.56	0.03190
3	6	8793.47	9248.22	8924.01	8716.79	8835.47	0.03182
3	7	8803.45	9316.85	8950.72	8716.59	8850.45	0.03175
3	8	8763.15	9334.87	8927.16	8666.45	8815.15	0.03185
4	1	8982.36	9144.08	9029.22	8956.36	8999.36	0.03136
4	2	8706.76	9044.31	8803.83	8650.34	8738.76	0.03222
4	**3**	**8651.97**	**9164.86**	**8799.25**	**8565.63**	**8698.97**	**0.03230**
4	4	8651.98	9340.42	8849.46	8535.51	8713.98	0.03218
4	5	8629.96	9493.74	8877.65	8483.55	8706.96	0.03214
4	6	8601.54	9640.53	8899.44	8425.34	8693.54	0.03213
4	7	8618.66	9833.11	8966.76	8412.40	8725.66	0.03195
4	8	8598.55	9988.32	8996.85	8362.38	8720.55	0.03191

Note. Model 1: Equal variances and covariances fixed to 0; Model 2: Varying variances and covariances fixed to 0; Model 3: Equal variances and equal covariances; Model 4: Varying variances and varying covariances; "Groups" indicates the number of Groups considered in the model; AIC = Akaike's Information Criterion; AWE = Approximate Weight of Evidence; BIC = Bayesian Information Criterion; CLC = Classification Likelihood Criterion; KIC = Kullback Information Criterion, C-RIV = Composite Relative Importance Vector. Based on an Analytic Hierarchy Process (AHP, see Akogul & Erisoglu, 2017) taking the mentioned fit indices into account, Model 4 with 3 groups (given in bold) showed overall the best fit (highest C-RIV).

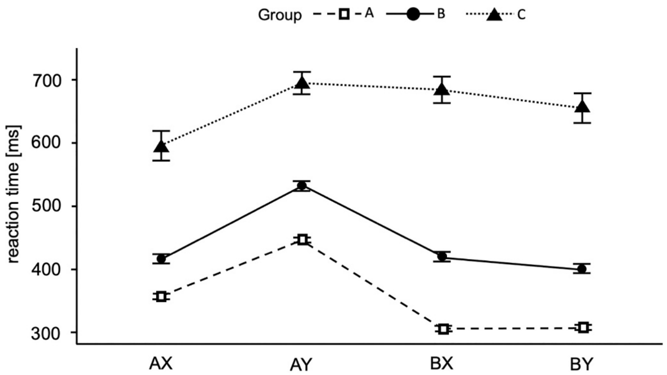


Fig. 2. Observed reaction time pattern and standard errors in the four conditions of the AX-CPT for the three groups identified by latent profile analysis.

difference that emerged for participants in Group A. This led to the conclusion that the difference for Group B between the BX and BY conditions, albeit significant, was so small that it could not be properly distinguished from the non-significant one that emerged for Group A. Summarized, Group B showed two RT differences that fit PMC, one that was RMC-consistent and some that did not comply with either. This implied that Group B engaged in a mix of PMC and RMC.

Group C participants had longer RTs in the BX condition than in the AX condition. Additionally, RTs in the BX condition were significantly longer than in the BY condition. These were both defined as markers for the engagement in RMC. For Group C the difference between the BX and BY condition was also significantly larger as the one that emerged in Group A, clearly distinguishing the two Groups.

Summarized, the RT pattern of Group A portrayed the expected PMC-consistent RT pattern while the RT pattern of Group C coincided with the RMC- consistent RT pattern. Although the interpretation of the RT pattern of Group B was less clear it seemed to have some similarities with PMC- and RMC-consistent patterns. It should also be mentioned that Group C was significantly slower in all AX-CPT conditions compared to the other two groups, and Group B was significantly slower than Group A.

7.3. *Identification of an item-position effect*

To analyse whether an IPE could be identified in the responses across the 18 items of the VMT, fixed-links modeling was applied. A congeneric model with the assumption of one underlying latent variable was compared to a bifactor model with a first latent variable representing reasoning ability and a second latent variable representing the IPE. Factor loadings on the latter were fixed with a quadratic increase to describe the increasing influence of the IPE from the first to the last item. The model fit statistics of the two models are given in Table 6.

According to Kenny (2015), the Comparative Fit Index (CFI) is seen as non-informative, when the RMSEA in the baseline model is lower than 0.158. The RMSEA of the baseline model was 0.112, therefore the CFI is listed in Table 6 but not used for model evaluation. The χ^2/df ratio was smaller than two for both models indicating good model fit (Wang & Wang, 2020). Also, according to SRMR and RMSEA, both models described the data well. However, the bifactor model had a lower AIC and BIC than the congeneric model indicating that it described the data better than the congeneric model. Additionally, in the bifactor model

Table 3
Estimates, t-values and *p*-values displayed for each Slope-as-Outcome model estimated with Group A as intercept and reaction time as dependent variable.

Model Aa			
Fixed effects	Estimate	t	*p*
Level 1			
Intercept (γ_{00})	356.87	67.14	<0.001
AY (γ_{10})	88.2	22.26	<0.001
BX (γ_{20})	−51.92	−13.11	<0.001
BY (γ_{30})	−50.11	−12.65	<0.001
Level 2			
Group B (γ_{01})	59.08	6.76	<0.001
Group C (γ_{02})	237.11	20.09	<0.001
AY:Group B (γ_{11})	27.67	4.25	<0.001
BX:Group B (γ_{21})	54.38	8.35	<0.001
BY:Group B (γ_{31})	34.22	5.26	<0.001
AY:Group C (γ_{12})	11.12	1.26	0.206
BX:Group C (γ_{22})	140.58	15.98	<0.001
BY:Group C (γ_{32})	110.08	12.51	<0.001

Model Ab			
Fixed effects	Estimate	t	*p*
Level 1			
Intercept (γ_{00})	445.07	83.73	<0.001
AX (γ_{10})	−88.2	−22.26	<0.001
BX (γ_{20})	−140.12	−35.37	<0.001
BY (γ_{30})	−138.31	−34.91	<0.001
Level 2			
Group B (γ_{01})	86.75	9.93	<0.001
Group C (γ_{02})	248.23	21.03	<0.001
AX:Group B (γ_{11})	−27.67	−4.25	<0.001
BX:Group B (γ_{21})	26.72	4.1	<0.001
BY:Group B (γ_{31})	6.56	1.01	0.314
AX:Group C (γ_{12})	−11.12	−1.26	0.206
BX:Group C (γ_{22})	129.46	14.72	<0.001
BY:Group C (γ_{32})	98.95	11.25	<0.001

Model Ac			
Fixed effects	Estimate	t	*p*
Level 1			
Intercept (γ_{00})	304.95	57.37	<0.001
AX (γ_{10})	51.92	13.11	<0.001
AY (γ_{20})	140.12	35.37	<0.001
BY (γ_{30})	1.81	0.46	0.648
Level 2			
Group B (γ_{01})	113.46	12.99	<0.001
Group C (γ_{02})	377.69	32	<0.001
AX:Group B (γ_{11})	−54.38	−8.35	<0.001
AY:Group B (γ_{21})	−26.72	−4.1	<0.001
BY:Group B (γ_{31})	−20.16	−3.1	0.002
AX:Group C (γ_{12})	−140.58	−15.98	<0.001
AY:Group C (γ_{22})	−129.46	−14.72	<0.001
BY:Group C (γ_{32})	−30.5	−3.47	<0.001

Note. Model Aa: Group A and condition AX as Intercept.
Model Ab: Group A and condition AY as Intercept.
Model Ac: Group A and condition BX as Intercept.
Bonferroni adjusted alpha value: 0.0011.

both the reasoning latent variable (φ = 0.211, $z = 7.034$, $p < .001$) as well as the latent variable representing the IPE explained a significant portion of variance (φ = 0.210, $z = 4.856$, $p < .001$). The reported variances were scaled as suggested by Schweizer and Troche (2019). The scaled variances clearly showed that the IPE and reasoning ability both explained an equal amount of variance in the bifactorial model, further emphasizing the relevance of considering the IPE as a latent variable in the measurement model.

Table 4
Estimates, t-values and *p*-values displayed for each Slope-as-Outcome model estimated with Group B as intercept and reaction time as dependent variable.

Model Ba			
Fixed effects	Estimate	t	*p*
Level 1			
Intercept (γ_{00})	415.95	59.99	<0.001
AY (γ_{10})	115.87	22.42	<0.001
BX (γ_{20})	2.46	0.48	0.633
BY (γ_{30})	−15.89	−3.07	0.002
Level 2			
Group C (γ_{01})	178.03	14.11	<0.001
Group A (γ_{02})	−59.08	−6.76	<0.001
AY:Group C (γ_{11})	−16.55	−1.76	0.078
BX:Group C (γ_{21})	86.19	9.17	<0.001
BY:Group C (γ_{31})	75.85	8.07	<0.001
AY:Group A (γ_{12})	−27.67	−4.25	<0.001
BX:Group A (γ_{22})	−54.38	−8.35	<0.001
BY:Group A (γ_{32})	−34.22	−5.26	<0.001

Model Bb			
Fixed effects	Estimate	t	*p*
Level 1			
Intercept (γ_{00})	531.82	76.7	<0.001
AX (γ_{10})	−115.87	−22.42	<0.001
BX (γ_{20})	−113.4	−21.95	<0.001
BY (γ_{30})	−131.75	−25.5	<0.001
Level 2			
Group C (γ_{01})	161.48	12.8	<0.001
Group A (γ_{02})	−86.75	−9.93	<0.001
AX:Group C (γ_{11})	16.55	1.76	0.078
BX:Group C (γ_{21})	102.74	10.93	<0.001
BY:Group C (γ_{31})	92.4	9.83	<0.001
AX:Group A (γ_{12})	27.67	4.25	<0.001
BX:Group A (γ_{22})	−26.72	−4.1	<0.001
BY:Group A (γ_{32})	−6.56	−1.01	0.314

Model Bc			
Fixed Effects	Estimate	t	*p*
Level 1			
Intercept (γ_{00})	418.41	60.35	<0.001
AX (γ_{10})	−2.46	−0.48	0.633
AY (γ_{20})	113.4	21.95	<0.001
BY (γ_{30})	−18.35	−3.55	<0.001
Level 2			
Group C (γ_{01})	264.22	20.95	<0.001
Group A (γ_{02})	−113.46	−12.99	<0.001
AX:Group C (γ_{11})	−86.19	−9.17	<0.001
AY:Group C (γ_{21})	−102.74	−10.93	<0.001
BY:Group C (γ_{31})	−10.34	−1.1	0.271
AX:Group A (γ_{12})	54.38	8.35	<0.001
AY:Group A (γ_{22})	26.72	4.1	<0.001
BY:Group A (γ_{32})	20.16	3.1	0.002

Note. Model Ba: Group B and condition AX as Intercept.
Model Bb: Group B and condition AY as Intercept.
Model Bc: Group B and condition BX as Intercept.
Bonferroni adjusted alpha value: 0.0011.

7.4. Relation of item position effect, reasoning and cognitive control

In a next step, factor scores for each participant were extracted from the bifactor model (see Fig. 3). To examine whether the three groups differed in their reasoning ability and/or in the extent of the IPE, six pairwise *t*-tests were calculated (see Table 7). To account for the multiple comparisons, alpha was adjusted to α = 0.0083. The reasoning factor scores did not differ between the three groups (see Table 7). Also, the IPE did not differ significantly between Group B and C nor between Group B and A. However, in Group A, the IPE was significantly more pronounced than in Group C. This difference was statistically significant even after alpha adjustment.

To be able to compare our results with previous research (e.g., Burgess & Braver, 2010; Gray et al., 2003), we also calculated pairwise *t*-tests for the VMT raw scores between the groups. Group A had significantly higher VMT scores when compared to Group C, $t(40) = -3.65$, $p < .002$, $d = -0.71$, while the VMT scores of Group B did not differ significantly from Group A, $t(138) = -1.295$, $p = .199$, $d = -0.19$, nor from Group C, $t(48) = 2.42$, $p = .019$, $d = 0.53$, when adjusting the alpha value for the three comparisons ($\alpha = 0.017$).

8. Discussion

In the present study, we found three clearly distinguishable groups of individuals by analysing RTs across the four conditions of the AX-CPT. Group A and Group B showed similar RT patterns, yet only the RT pattern of Group A directly coincided unambiguously with the one assumed for PMC indicating strong engagement in PMC. Group B exhibited mixed engagement in PMC and RMC. Group C had an RT pattern that resembled the one expected for RMC. Across all

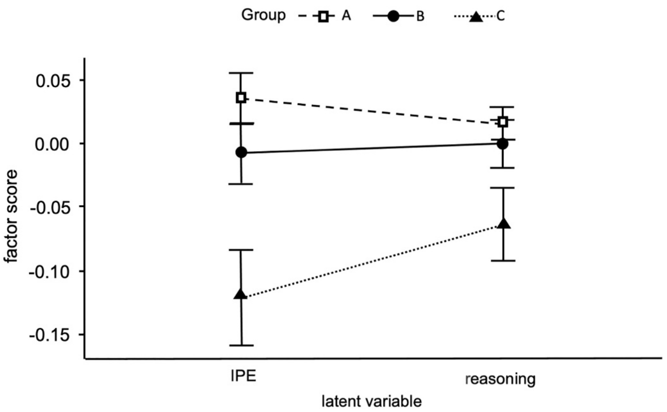


Fig. 3. Factor scores on the latent variables representing the item-position-effect (IPE) and reasoning ability in the Vienna Matrices Test for the three groups identified. Error bars represent standard errors.

Table 5
Estimates, t-values and *p*-values displayed for each Slope-as-Outcome model estimated with Group C as intercept and reaction time as dependent variable.

Model Ca

Fixed effects	Estimate	t	*p*
Level 1			
Intercept (γ_{00})	593.98	56.36	<0.001
AY (γ_{10})	99.32	12.64	<0.001
BX (γ_{20})	88.66	11.29	<0.001
BY (γ_{30})	59.97	7.63	<0.001
Level 2			
Group B (γ_{01})	−178.03	−14.11	<0.001
Group A (γ_{02})	−237.11	−20.09	<0.001
AY:Group B (γ_{11})	16.55	1.76	0.078
BX:Group B (γ_{21})	−86.19	−9.17	<0.001
BY:Group B (γ_{31})	−75.85	−8.07	<0.001
AY:Group A (γ_{12})	−11.12	−1.26	0.206
BX:Group A (γ_{22})	−140.58	−15.98	<0.001
BY:Group A (γ_{32})	−110.08	−12.51	<0.001

Model Cb

Fixed effects	Estimate	t	*p*
Level 1			
Intercept (γ_{00})	693.3	65.79	<0.001
AX (γ_{10})	−99.32	−12.64	<0.001
BX (γ_{20})	−10.66	−1.36	0.175
BY (γ_{30})	−39.35	−5.01	<0.001
Level 2			
Group B (γ_{01})	−161.48	−12.8	<0.001
Group A (γ_{02})	−248.23	−21.03	<0.001
AX:Group B (γ_{11})	−16.55	−1.76	0.078
BX:Group B (γ_{21})	−102.74	−10.93	<0.001
BY:Group B (γ_{31})	−92.4	−9.83	<0.001
AX:Group A (γ_{12})	11.12	1.26	0.206
BX:Group A (γ_{22})	−129.46	−14.72	<0.001
BY:Group A (γ_{32})	−98.95	−11.25	<0.001

Model Cc

Fixed effects	Estimate	t	*p*
Level 1			
Intercept (γ_{00})	682.63	64.77	<0.001
AX (γ_{10})	−88.66	−11.29	<0.001
AY (γ_{20})	10.66	1.36	0.175
BY (γ_{30})	−28.69	−3.65	<0.001
Level 2			
Group B (γ_{01})	−264.22	−20.95	<0.001
Group A (γ_{02})	−377.69	−32	<0.001
AX:Group B (γ_{11})	86.19	9.17	<0.001
AY:Group B (γ_{21})	102.74	10.93	<0.001
BY:Group B (γ_{31})	10.34	1.1	0.271
AX:Group A (γ_{12})	140.58	15.98	<0.001
AY:Group A (γ_{22})	129.46	14.72	<0.001
BY:Group A (γ_{32})	30.5	3.47	<0.001

Note. Model Ca: Group C and condition AX as Intercept.
Model Cb: Group C and condition AY as Intercept.
Model Cc: Group C and condition BX as Intercept.
Bonferroni adjusted alpha value: 0.0011.

Table 7
Results for two-tailed *t* tests to compare factor scores for the latent variables reflecting the item-position effect (IPE) and reasoning ability between the three groups with different engagement in PMC and RMC.

	t	df	*p*	*Cohen's d*
IPE				
Group B – C	2.55	51.23	.014	0.57
Group B – A	−1.44	148.55	.151	0.22
Group C – A	3.72	44.94	.00056	0.76
Reasoning				
Group B – C	1.84	51.76	.071	0.41
Group B – A	−0.73	129.62	.471	0.11
Group C – A	2.51	40.14	.016	0.54

Note. Bonferroni adjusted alpha value: 0.0083.

Table 6
Chi-square (degrees of freedom), *p*-value, and various fit indices to compare the congeneric model and the bifactorial model which includes latent variables representing reasoning ability and item-position effect.

Model	χ^2 (*df*)	*p*	RMSEA	SRMR	CFI	AIC	BIC
Congeneric	217.28 (135)	<.001	0.054	0.064	0.794	2999.44	3119.93
Bifactorial	208.99 (134)	<.001	0.052	0.063	0.812	2993.15	3116.99

Note. Root Mean Squared Error Approximation (RMSEA); Standardized Root Mean Square Residual (SRMR); Akaike Information Criterion (AIC); Bayesian Information Criterion (BIC); Comparative Fit Index (CFI) is non-informative as the RMSEA of the baseline model is lower than 0.158.

participants, we identified an IPE in the VMT data indicating that individuals differed in the extent they could benefit from the completion of previous items during the completion of later items. Although the effects were partly of medium size, the three groups did not differ significantly in their reasoning ability. The IPE, however, was more pronounced in Group A compared to Group C, which is in line with the assumption, that engagement in PMC is associated with a larger IPE.

8.1. Classification of groups

The LPA on RTs in the four conditions of the AX-CPT identified three groups of individuals. Results of the MLM led to the following characterization of the groups: Participants in Group A had shorter RTs than the other two groups and the RT pattern was a straightforward match with the expected pattern for individuals applying PMC. Participants in Group B had somewhat slower RTs than Group A. Unlike Group A, their fastest RTs were in the AX, BX and the BY condition and there was a significant, albeit very small difference between the BX and BY condition. The RT pattern of Group B showed features typical for PMC as well as features typical for RMC. Therefore, we interpreted the RT pattern of Group A as engagement in PMC, and the RT pattern of Group B as a mixture of engagement in PMC and RMC.

Group C had not only slower RTs compared to the other two groups but also a very different RT pattern. The difference between RTs in the AY and in the BX condition, which has been previously emphasized as an indicator for using PMC (Braver et al., 2001), was significantly smaller in Group C than in the other groups, indicating that Group C engaged less in PMC than the other two groups. Also, the expected difference between RTs in the BX and BY condition was significantly larger in Group C compared to Group A. The emergence of a difference between the BX and BY condition is a clear marker for engagement in RMC. A further marker would have been similar RTs in the AY and BY conditions. This was not the case for any group. Yet the differences between the conditions were notably larger for Groups A and B, and significantly smaller for Group C. This difference between the AY and BY conditions, albeit significant, is very small for Group C. Further support for the assumption that Group C most likely engaged in RMC can be taken from the findings reported by Gonthier, Macnamara, et al. (2016). The authors explicitly manipulated the AX-CPT to make individuals engage more strongly in RMC. The RT pattern which resulted from this manipulation was similar to the RT pattern observed in the present study for Group C with longer RTs in the AY condition than in the other three conditions. The above-mentioned difference between RTs in the AY condition compared to the BX and BY conditions was even more pronounced in the study by Gonthier, Macnamara, et al. (2016) than in our Group C.

In sum, three clearly distinguishable groups could be identified in the present study, which did not only differ in overall RT or their RTs in single conditions, but in their RT patterns across the four AX-CPT conditions. The RT pattern of Group A clearly matched the assumed pattern for PMC, the RT pattern of Group B indicated mixed engagement in PMC and RMC while the RT pattern of Group C indicated engagement in RMC.

8.2. Relation of item position effect, reasoning and cognitive control

When examining the association between reasoning ability and the two mechanisms of cognitive control, previous studies (Burgess & Braver, 2010; Gray et al., 2003) split the sample of participants into subgroups according to their reasoning ability score and declared the groups as high and low Gf individuals. Then behavioural data and/or neural activity between these subgroups were compared regarding their engagement in PMC/RMC. The results of these previous studies suggested that high Gf individuals engaged more strongly in PMC than low Gf individuals. Results were seen as evidence for the idea that the larger cognitive resources of individuals with high Gf facilitated the use of the resource-demanding PMC (Braver, 2012). When we directly compared the VMT raw scores between the three groups identified in the present study, we obtained similar results: Group A, which most strongly engaged in PMC, had significantly higher reasoning scores (as indicator of Gf) compared to Group C which had the weakest or no engagement in PMC and showed strong evidence for using RMC. This is worth to mention since we used the VMT in the present study to measure reasoning ability while Burgess and Braver (2010) as well as Gray et al. (2003) used Raven's APM. Thus, the outcome of a functional relationship between reasoning ability as a measure of Gf and the dual mechanisms of cognitive control seems not to depend on the instrument with which reasoning ability is assessed.

In contrast to previous studies, however, we extracted an IPE from the present reasoning test. The existence of an IPE in reasoning test data in addition to a latent variable representing reasoning ability was in line with an increasing body of research on the IPE in reasoning measures (Ren et al., 2014; Ren et al., 2015; Sun et al., 2019; Troche et al., 2016; Zeller et al., 2017). Both latent variables explained an equal proportion of variance in the measurement model indicating that the IPE cannot be neglected when reasoning ability is correlated with other variables. To date, the most plausible explanation for the IPE states that some individuals strongly benefit from already completed items, while others do not (Ren et al., 2014). Therefore, some individuals are better at using knowledge gained during the completion of earlier items to ideally bias their information processing for the completion of later items. Proceeding from this interpretation of the IPE, we assumed that individuals using PMC showed a larger IPE than individuals using RMC due to their early selection and maintenance of (context) information to bias attention in an ideal manner during the completion of the task at hand (cf. Braver, 2012). This idea was supported by our empirical results as individuals who strongly engage in PMC (Group A) exhibited a more pronounced IPE compared to individuals who engage in RMC (Group C), while the groups did not statistically differ in their reasoning ability. This result is remarkable as it suggests that the direction of the relationship between the engagement in RMC/PMC and fluid intelligence might be interpreted differently than previously proposed by Braver and his colleagues (Burgess & Braver, 2010; Gray et al., 2003). These authors argued that higher Gf as a reflection of higher cognitive capacities facilitates applying the resource demanding PMC. On the contrary, our results suggest that using PMC rather than RMC leads to higher reasoning test scores because of a more adaptive behaviour during test completion. Individuals engaging in PMC seem to use context information, knowledge gained from solving previous items, to solve later items. This leads to a stronger IPE, while individuals engaging in RMC seem to benefit less from previously solved items and therefor have a smaller IPE. It is important to mention that, although Group A and Group C did not differ significantly in their reasoning ability, the effect size was quite large with Cohen's $d = 0.54$ so that the size of Group C was perhaps not large enough to reveal a significant difference in reasoning ability when compared to the other groups. A more tentative interpretation, therefore holds, that Group A and C differed primarily in their IPE and only subordinately in their reasoning ability. The differences between Group B and Group C in the IPE and reasoning ability might be similarly interpreted against the obtained effect sizes presented in Table 7.

8.3. Limitations

From this point of view, the rather small size of Group C might be considered a limitation of the present study since it resulted in larger standard errors when compared to the other two groups. In contrast, more than half of the sample belonged to Group A showing a typical PMC pattern. This was surprising as we composed the sample not only from university students but also from individuals without university entrance certification to spread the range of intelligence. As a result, the IQ distribution in our sample was highly similar to the distribution in the norm sample. Nevertheless, the portion of individuals identified as applying RMC was rather small. Therefore, it could be interesting to see whether a similar group classification would be obtained in a sample with a larger age range or in a sample of older adults for whom Braver et al. (2001) reported more engagement in RMC compared to younger individuals. Additionally, a combination of a classification approach as introduced in the present study based on behavioural data and a neurophysiological approach using fMRI might illuminate whether the groups would show brain activation patterns reported to be PMC- or RMC-specific (Braver et al., 2009; Paxton et al., 2008).

Since the experimental approach as well as the statistical methods of our study do not allow for a strong causal interpretation of the relationship between PMC and IPE, a more straightforward hypothesis test is called for to further confirm our interpretation. Nevertheless, our results suggest that another causal relationship might be conceivable between measures of Gf and the dual mechanisms of cognitive control than suggested by Burgess and Braver (2010).

9. Conclusion

To summarize, three clearly distinguishable groups could be identified, which differed in their engagement in PMC and RMC and in their VMT test scores. Albeit, under the consideration of the IPE in the VMT data, the identified groups did not differ in their reasoning ability. Instead, the difference in the test scores could be explained by a more pronounced IPE in the group with strong engagement in PMC compared to the group that engaged in RMC. These results present first evidence for the notion that using PMC rather than RMC can lead to better reasoning test scores due to a stronger IPE. In other words, compared to individuals who engage in RMC, individuals engaging strongly in PMC benefit more from solving previous items when they solve later items in a reasoning test.

CRediT authorship contribution statement

Helene M. von Gugelberg: Conceptualization, Methodology, Data curation, Formal analysis, Visualization; Writing - original draft, Writing - review & editing; **Karl Schweizer**: Conceptualization, Validation, Methodology, Writing - review & editing; **Stefan J. Troche**: Conceptualization, Methodology, Supervision, Writing - review & editing, Resources, Project administration.

References

Akogul, S., & Erisoglu, M. (2017). An approach for determining the number of clusters in a model-based cluster analysis. *Entropy, 19*(9), 452. https://doi.org/10.3390/e19090452

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences, 16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Braver, T. S., Barch, D. M., Keys, B. A., Carter, C. S., Cohen, J. D., Kaye, J. A., Janowsky, J. S., Taylor, S. F., Yesavage, J. A., Mumenthaler, M. S., Jagust, W. J., & Reed, B. R. (2001). Context processing in older adults: Evidence for a theory relating cognitive control to neurobiology in healthy aging. *Journal of Experimental Psychology: General, 130*(4), 746–763. https://doi.org/10.1037/0096-3445.130.4.746

Braver, T. S., Paxton, J. L., Locke, H. S., & Barch, D. M. (2009). Flexible neural mechanisms of cognitive control within human prefrontal cortex. *Proceedings of the National Academy of Sciences, 106*(18), 7351–7356. https://doi.org/10.1073/pnas.0808187106

Burgess, G. C., & Braver, T. S. (2010). Neural mechanisms of interference control in working memory: Effects of interference expectancy and fluid intelligence. *PLoS One, 5*(9), Article e12861. https://doi.org/10.1371/journal.pone.0012861

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge University Press.

Der, G., & Deary, I. J. (2017). The relationship between intelligence and reaction time varies with age: Results from three representative narrow-age age cohorts at 30, 50 and 69 years. *Intelligence, 64*, 89–97. https://doi.org/10.1016/j.intell.2017.08.001

DiStefano, C. (2016). Examining fit with structural equation modeling. In K. Schweizer, & C. DiStefano (Eds.), *Principles and methods of test construction* (pp. 166–196). Hogrefe.

Formann, A. K., Piswanger, K., & Waldherr, K. (2011). *Wiener Matrizen-Test 2: Ein Rasch-skaldierter sprachfreier Kurztest zu Erfassung der Intelligenz.* Hogrefe.

Gonthier, C., Braver, T. S., & Bugg, J. M. (2016). Dissociating proactive and reactive control in the stroop task. *Memory & Cognition, 44*(5), 778–788. https://doi.org/10.3758/s13421-016-0591-1

Gonthier, C., Macnamara, B. N., Chow, M., Conway, A. R. A., & Braver, T. S. (2016). Inducing proactive control shifts in the AX-CPT. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01822

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6*(3), 316–322. https://doi.org/10.1038/nn1014

Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*(3), 179–203. https://doi.org/10.1016/0160-2896(84)90008-4

Jensen, A. R. (1998). The g factor and the design of education. In R. J. Sternberg, & W. M. Williams (Eds.), *Intelligence, instruction, and assessment: Theory into practice* (pp. 111–131). Lawrence Erlbaum Associates Publishers.

Kan, K.-J., Kievit, R. A., Dolan, C., & der Maas, H.v. (2011). On the interpretation of the CHC factor Gc. *Intelligence, 39*(5), 292–302. https://doi.org/10.1016/j.intell.2011.05.003

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods, 39*, 979–984. https://doi.org/10.3758/BF03192993

Kenny, D. A. (2015). *Measuring model fit*.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13). https://doi.org/10.18637/jss.v082.i13

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the kenward-Roger correction. *Multivariate Behavioral Research, 52*(5), 661–670. https://doi.org/10.1080/00273171.2017.1344538

Paxton, J. L., Barch, D. M., Racine, C. A., & Braver, T. S. (2008). Cognitive control, goal maintenance, and prefrontal function in healthy aging. *Cerebral Cortex, 18*(5), 1010–1028. https://doi.org/10.1093/cercor/bhm135

Raven, J., & Raven, J. (2003). Raven progressive matrices. In *Handbook of nonverbal assessment* (pp. 223–237). Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-1-4615-0153-4_11.

Redick, T. S. (2014). Cognitive control in context: Working memory capacity and proactive control. *Acta Psychologica, 145*, 1–9. https://doi.org/10.1016/j.actpsy.2013.10.010

Ren, X., Schweizer, K., Wang, T., Chu, P., & Gong, Q. (2017). On the relationship between executive functions of working memory and components derived from fluid intelligence measures. *Acta Psychologica, 180*, 79–87. https://doi.org/10.1016/j.actpsy.2017.09.002

Ren, X., Schweizer, K., Wang, T., & Xu, F. (2015). The prediction of students' academic performance with fluid intelligence in giving special consideration to the contribution of learning. *Advances in Cognitive Psychology, 11*(3), 97–105. https://doi.org/10.5709/acp-0175-z

Ren, X., Wang, T., Altmeyer, M., & Schweizer, K. (2014). A learning-based account of fluid intelligence from the perspective of the position effect. *Learning and Individual Differences, 31*, 30–35. https://doi.org/10.1016/j.lindif.2014.01.002

Rosenberg, J. M., Beymer, P. N., Anderson, D. J., Van Lissa, C. J., & Schmidt, J. A. (2019). tidyLPA: An R package to easily carry out latent profile analysis (LPA) using open-source or commercial software. *Journal of Open Source Software, 3*(30), 978. https://doi.org/10.21105/joss.00978

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schweizer, K. (2013). A threshold-free approach to the study of the structure of binary data. *International Journal of Statistics and Probability, 2*(2), 67. https://doi.org/10.5539/ijsp.v2n2p67

Schweizer, K., & Troche, S. (2018). Is the factor observed in investigations on the item-position effect actually the difficulty factor? *Educational and Psychological Measurement, 78*(1), 46–69. https://doi.org/10.1177/0013164416670711

Schweizer, K., & Troche, S. (2019). The EV scaling method for variances of latent variables. *Methodology, 15*(4), 175–184. https://doi.org/10.1027/1614-2241/a000179

Schweizer, K., Troche, S. J., & Rammsayer, T. H. (2011). On the special relationship between fluid and general intelligence: New evidence obtained by considering the position effect. *Personality and Individual Differences, 50*(8), 1249–1254. https://doi.org/10.1016/j.paid.2011.02.019

Sodian, B., & Frith, U. (2008). Metacognition, theory of mind, and self-control: The relevance of high-level cognitive processes in development, neuroscience, and education. *Mind, Brain, and Education, 2*(3), 111–113. https://doi.org/10.1111/j.1751-228X.2008.00040.x

Sun, S., Schweizer, K., & Ren, X. (2019). Item-position effect in Raven's matrices: A developmental perspective. *Journal of Cognition and Development, 20*(3), 370–379. https://doi.org/10.1080/15248372.2019.1581205

Troche, S. J., Wagner, F. L., Schweizer, K., & Rammsayer, T. H. (2016). The structural validity of the culture fair test under consideration of the item-position effect. *European Journal of Psychological Assessment, 35*(2), 182–189. https://doi.org/10.1027/1015-5759/a000384

Wang, J., & Wang, X. (2020). *Structural equation modeling - Applications using Mplus* (2nd ed.). Wiley.

Zeller, F., Wang, T., Reiß, S., & Schweizer, K. (2017). Does the modality of measures influence the relationship among working memory, learning and fluid intelligence? *Personality and Individual Differences, 105*, 275–279. https://doi.org/10.1016/j.paid.2016.10.013

Appendix B

Study 2: Strategy and the Learning Hypothesis

Individual Differences in Strategy and the Item-Position Effect in Reasoning Ability Measures.

Helene M. von Gugelberg & Stefan J. Troche

Department of Psychology, University of Bern, Switzerland

Author Note:

Helene M. von Gugelberg https://orcid.org/0000-0002-7971-5038
Stefan J. Troche https://orcid.org/0000-0002-0961-1081

Correspondence concerning this article should be addressed to Helene M. von Gugelberg, Department of Psychology, University of Bern, Switzerland. Email: helene.vongugelberg@unibe.ch

Abstract

Despite the high similarity of reasoning ability items, research indicates that individuals apply different strategies when solving them (Bethell-Fox et al., 1984). The two distinct strategies are response elimination and constructive matching. The latter frequently showing a positive correlation with reasoning ability (e.g., Vigneau et al., 2006) entails the individual systematically scanning the presented problem matrix of an item, before scanning the response alternatives. To further the investigation of what lies at the source of individuals applying different strategies during test taking, a study tracking eye movement during the solving process of the Advanced Progressive Raven Matrices (APM) was conducted. Results showed in line with other research (e.g., Vigneau et al., 2006) a positive correlation of reasoning ability and constructive matching. Results further indicated that participants used more constructive matching towards the end of the APM. This is the opposite of what Bethel-Fox et al. (1984) or Gonthier and Roulin (2020) observed in their work. This change in strategy was correlated with the item-position effect detected in the APM scores. The item-position effect captures the increasing score variance in a test with homogenous item materials such as the APM and is assumed to relate to rule learning (Ren et al., 2015). Possible reasons for the diverging results and the newfound relation to the item-position effect are discussed.

Fluid intelligence as defined by Carroll (1993) includes the abilities to solve logical puzzles, abstract problems or to infer rules in a set of figures. These abilities are often referred to as reasoning abilities and are strong predictors for an abundance of outcomes (e.g., Gottfredson & Deary, 2004). The Advanced Raven Progressive Matrices (APM, Raven, Raven, & Court 1998) are a well-known reasoning ability measure with homogenous items. Each problem or item the individual has to solve, is a 3 × 3 matrix that depicts a certain pattern of geometrical shapes with the bottom right entry missing. For each item, the individual selects one out of eight response alternatives to complete the pattern in the matrix (see Figure 1).

Despite the high similarity of items, people have different ways of going about solving such items measuring reasoning ability (Bethell-Fox et al., 1984). Just because two individuals achieved the same score on a test, it does not necessarily mean, that they used the same abilities or strategies to get there. What abilities are truly at play when solving tests such as the APM is crucial knowledge to further understand human intelligence and apparently, “individuals not only do things differently when asked to solve intelligence test items: they also do different things” (p. 271, Vigneau et al., 2006), since individuals seem to vary in the strategy applied during test taking.

Two distinct strategies were detected through verbal protocols and observation of eye movements (Snow, 1978), while individuals were solving reasoning ability measures that relied on the common multiple-choice format such as the APM. The constructive matching strategy describes the fact that individuals spend a lot of time on the matrix, identifying the different rules corresponding to the pattern. From the identified rules the individual then mentally constructs the missing entry and selects the matching entry from the response alternatives. Hence the name, constructive matching. The other dominant strategy is response elimination. Here the individual eliminates step by step non-viable solutions from the response alternatives that are presented, therefore eliminating unsuitable responses until a solution is found.

These early findings of Snow (1978) were replicated in eye tracking data (Vigneau et al., 2006) and verbal protocols (Jarosz et al., 2019) but also with questionnaires (Gonthier & Thomassin, 2015). Regardless of the way strategy was assessed the majority of studies confirmed (e.g., Gonthier & Roulin, 2020; Jastrzebski et al., 2018) that constructive matching was associated with higher test scores compared to response elimination.

From the mentioned strategy measures, eye tracking data is an objective measure used in several studies (Vigneau et al., 2006; Hayes et al., 2015; Laurence et al., 2018) and offers information on every item for every individual. Within eye racking data, the toggle rate is by far the most intuitive and straightforward indicator of strategy use. The toggle rate is based on fixations an individual makes while solving an item, i.e. the time intervals when the eyes are fixed to a given area. A toggle occurs when an individual first shows a fixation on the matrix, followed by a fixation on the response alternatives, or vice versa.

Constructive matching leads to very few toggles since most of the time is spent on the matrix to analyze the rules and then construct the solution. Individuals would only alternate a few times from looking at the matrix to the response alternatives before selecting an answer. With response elimination it is necessary to alternate very often since the solution is found through the process of elimination. This results in a lot of toggles. Thus, constructive matching and response elimination differ in the frequency of toggles during the processing of a test item. Laurence et al. (2018) reported that individual differences in the toggle rate explained up to 45% of the variance in reasoning ability.

Yet the reasons why different strategies are used and why they lead to different results is an ongoing debate. Contemporary research indicates that on the one hand, test properties (e.g., Raden & Jarosz, 2022) and on the other hand, individual differences in mental resources (Gonthier & Thomassin, 2015; Jarosz et al., 2019; Li et al., 2022) influence what strategy an individual applies to an item. Summarized, strategy use is influenced by the interrelationship of mental resources of an individual and perceived item properties. Bethell-Fox et al. (1984) and Snow (1980) came to the same conclusions in their pioneering work. They found when the capacity of an individual to hold rules and manipulate objects in the mind is exceeded, they switch from constructive matching to response elimination as a fallback strategy.

Such a shift in strategy was observed in the original version of the APM by Gonthier and Roulin (2020). Their data showed that all participants progressively shifted from constructive matching towards response elimination. Therefore, the current state of research suggests that individuals not only “do different things” in terms of strategy use as suggested by Vigneau et al. (2006), but also adapt their strategy use individually to perceived item properties throughout a test. These individual differences, in how people engage in strategy throughout a test, could lead to an increase of variance throughout a test.

These individual differences can create a systematic change from the first to the last item in a reasoning test and has also been described in studies on the structure of reasoning tests (Schweizer, 2006). More specifically, a series of studies used confirmatory factor analysis with a fixed-links modeling approach where two latent variables are extracted from the items of reasoning tests (e.g., Ren et al., 2015). In such a model one latent variable describes reasoning ability and an additional latent variable from the same series of items describes increasing individual differences across the test. To account for the increasing individual differences the factor loadings of said latent variable are fixed to monotonically increase from the first to the last item. Due to the close relationship between the items position within the test and the corresponding factor loading, this latent variable is referred to as the item-position effect.

Including the item-position effect in measurement models on reasoning tests has repeatedly shown to improve their structural description. Therefore, indicating a source of variance in the data that increases from the first to the last item (e.g., Schweizer et al., 2012; Troche et al., 2016). Yet, the very nature of this source of variance is still unclear.

Previous research ruled out item difficulty since the item-position effect also occurred when items were not ordered according to their difficulty (Schweizer et al., 2021; Zeller et al., 2017) or when item difficulty was statistically controlled for (Schweizer & Troche, 2018). Other studies showed that the item-position effect was related to working memory updating and shifting (Ren et al., 2017) as well as to proactive control (von Gugelberg et al., 2021) and rule learning (Ren et al., 2014; Schweizer et al., 2021).

Even with the ongoing debate about the origin of the item-position effect, it should not be neglected. The mere fact, that it consistently can be detected in measurement models of reasoning ability measures speaks for its existence and underlines the necessity to account for it. If simply not to draw false conclusions due to subpar data description. This leads to the objectives of the current study.

Current Study

From the current state of research, it can be concluded that constructive matching is clearly positively related to performance in the APM (e.g., Jarosz et al., 2019; Jastrzebski et al., 2018). Additionally, Gonthier & Roulin (2020) found in their data based on questionnaires, that participants changed their strategy throughout test completion while Vigneau et al., (2006) found no shift in strategy, but rather an initial difference between subjects on what strategy they engage in.

Both studies point toward individual differences in the applied strategy during test completion. To account for this idiosyncratic behavior during test completion, a confirmatory factor analysis with the fixed-links approach is tested on the eye tracking data (i.e., toggle rate). This statistical approach allows for potential individual differences in the overall strategy applied (as in Vigneau et al., 2006) but also for a shift in strategy (as in Gonthier & Roulin, 2020). Therefore, the first objective of the current study is to analyze whether these individual differences occur and can be depicted in a bifactor model based on the toggle rate.

The second objective of the current study is to analyze whether an item-position effect can be identified in the APM scores. If a more adequate data description includes the item-position effect, it should be included in any further analysis.

The third objective of the current study is to analyze whether reasoning ability is related to the toggle rate, and if individual differences in toggle rate throughout test completion can be identified, whether it is related to reasoning ability and / or the item-position effect.

Method

Participants

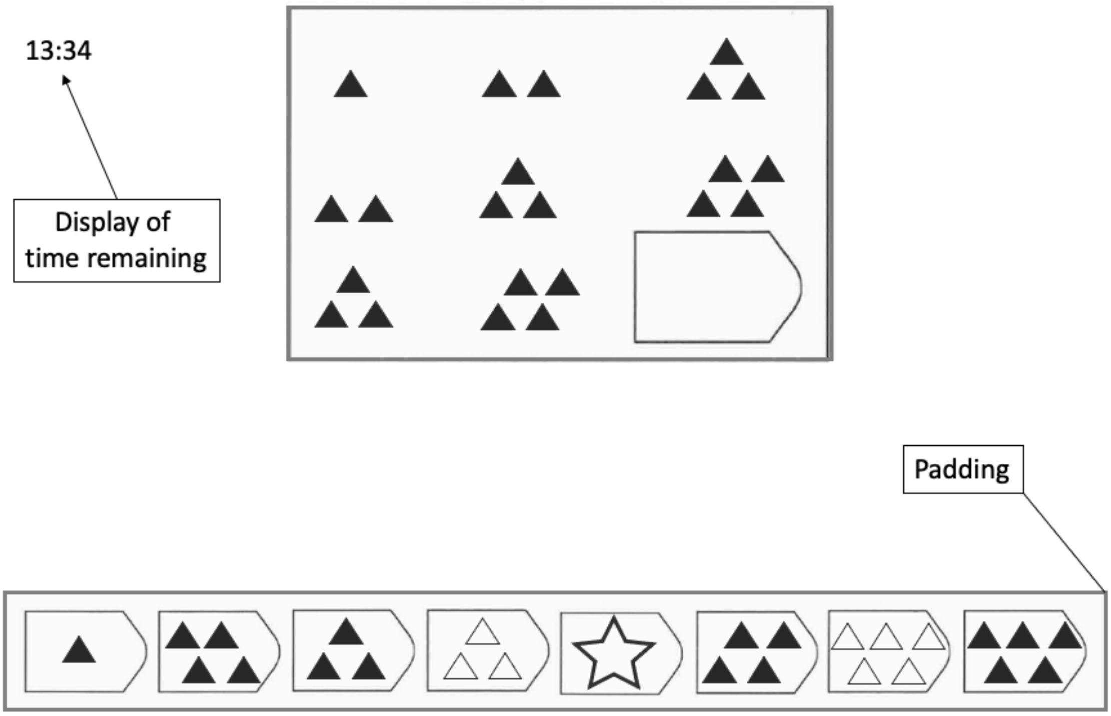
Participants were recruited through the university, other local institutions, online or directly by the contributors of the project. Psychology undergraduates received course credit and participants without a university entrance qualification received 20 CHF for participating. A total of 217 individuals participated. One participant did not declare gender, age nor highest level of education. The rest of the sample had a mean age of 27.53 years (SD = 11.91 years). Thereof 136 participants described themselves as female, 79 as male and one person chose “other”. 136 participants reported having university entrance qualification as their highest level of education, 53 a bachelor’s degree or higher, and 27 participants had neither. All participants reported normal or corrected-to-normal vision and gave written informed consent. The study protocol was approved by the local ethics committee of the University of Bern (No. 2020-07-00001).

Raven’s Advanced Progressive Matrices

The APM is aimed at an high aptitude population (Raven et al., 1998). The black and white stimuli of the APM each consist of 3 × 3 problem matrix with eight geometrical figures and the bottom right entry missing. Underneath the problem matrix eight response alternatives are presented. As instructed by the manual, participants completed two example items followed by the 36 items given in the predetermined order. Contrary to the manual, a time limit was set at 30 minutes for the completion of the 36 items. Additionally, for the ease of gathering eye tracking data, all response alternatives were placed in a line (see Figure 1). The APM score (1 = correct / 0 = false) for each item was used for the analysis.

Figure 1

Fictious example item of the APM



Toggle Rate

For the analysis of eye movement data, monocular eye data as used, whereby the eye with smaller measurement error was automatically selected by the EyeLink (SR Research, 2016) system after the calibration and validation procedure. Fixation duration threshold was set at 100 ms (e.g., Martarelli & Mast, 2013). For the area of response alternatives and also the matrix, an interest area was created. The interest area for the response alternatives had additional padding (Bojko, 2013). No padding was added to the interest area of the problem matrix, since it already contains blank space around the matrix entries (see Figure 1).

Data of each individual was checked for drift. Drift occurs when a participant moves their head after the calibration and validation process. This leads to a systematic drift in a certain direction, leaving many or even all fixations on blank areas of the screen. Whenever such a systematic drift was detected, all fixations were moved in cohesion to have them located on reasonable areas of the stimuli. After the raw data inspection, a fixation report for each participant was created and the number of toggles was calculated. A toggle was defined as a fixation on the matrix followed by a fixation on the response alternatives or vice versa. All fixations outside of the defined interest areas were recoded with the value of the interest area the previous fixation was in. Therefore, if a participant first fixates on the matrix, then stares at empty white space outside of the interest areas (possibly thinking) and then shows the next fixation on the response alternatives, this would be counted as a toggle. Without the recoding, this type of scan path would not be counted as a toggle, although it qualifies as an alternation between matrix and response alternatives during the solving process. For each item then the number of toggles was divided though time spent on the corresponding item (item latency). Toggle rate of each item was used for the analysis.

Apparatus

The APM was implemented with the Psychopy v2020.1.2 (Peirce et al., 2019) software. Participants used a wired computer mouse to select their answers. Eye data collection was made with the the Eyelink 1000 Plus system (SR Research, 2016). Our Eyelink setup, had participants use a chin-forehead rest and tracked eye movements with an infrared video camera with a 500hz sampling rate. Participants solved the APM on an 18-inch *Dell* computer screen with a resolution of 1280 x 1024 pixels seated 850 mm in front of it. Eye movement data as processed using Eyelink's Data Viewer Software (SR Research, 2016).

Procedure

Data was collected as part of a larger two-part study. During the first session, participants completed several tasks not relevant for the present study and also answered a socio-demographic questionnaire. The second session took place at least 24 hours after the first session. During the second session eye movement data and other physiological data was collected during the completion of several tasks. After attaching all electrodes and instructing participants about the chin-forehead rest, the data collection of the second session started with participants completing the APM.

The start of the APM initialized the Eyelink 1000 Plus system (SR Research, 2016), and prompted the experimenter to calibrate the eye tracking measurements. After successful calibration, a validation was completed. Standard 9-point calibration and validation procedure was applied until an eye-tracking error below 0.8° was obtained. After reading the instructions of the APM and solving two example items, participants had 30 minutes to solve all the APM items. The remaining time was displayed in the top left corner of the screen. Each item was separated by a stimulus interval of 2 seconds. During this interval a fixation cross was presented in the middle of the screen. When the time limit was exceeded or all 36 items were answered, participants completed other tasks, not relevant to the present study.

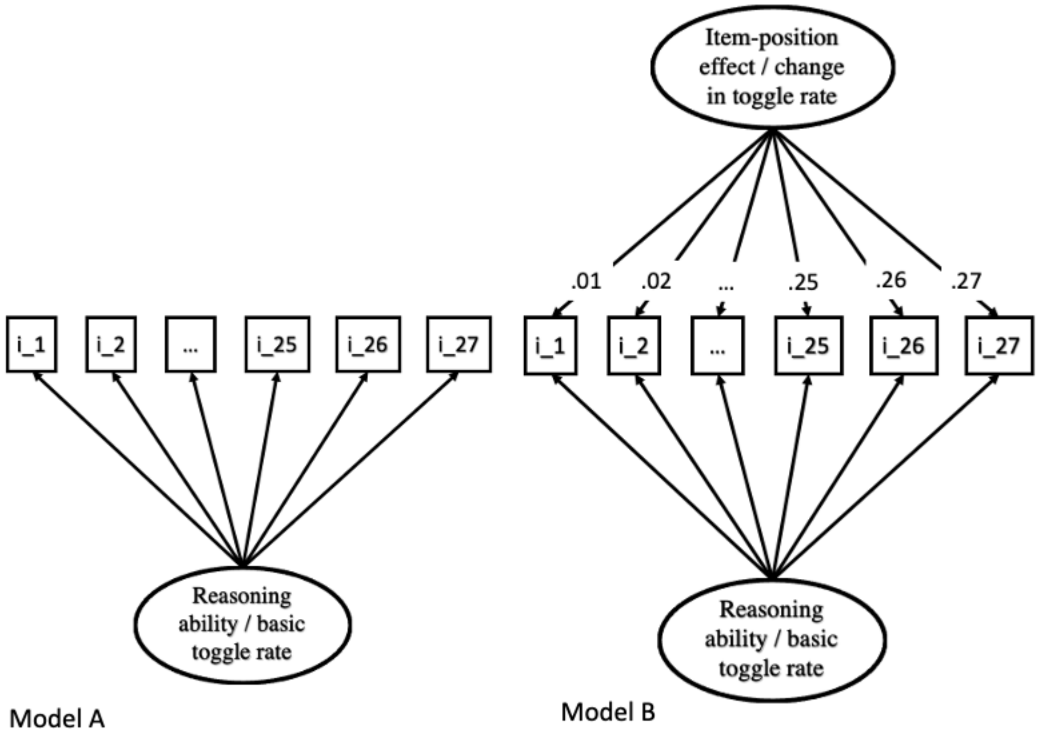
Statistical Analysis

For the Analyses the R packages *lavaan* (Rosseel, 2012) and *psych* (Revelle, 2011) were used. All confirmatory factor analyses were run with robust maximum likelihood estimation.

For the first objective of this study, two different models were fit to the toggle rate data. The goal was to analyse whether a bifactor model (Figure 2, Panel B) indeed can describe the toggle rate and its proposed variability more accurately than a one-factor model (Figure 2, Panel A). For the bifactor model, a second latent variable was introduced. For this second latent variable, the factor loadings were set to linearly increase from the first to the last item (Schweizer & Troche, 2018). This second latent variable would depict a change in toggle rate throughout the test and possibly reflects a change in strategy, that occurs throughout test taking (Gonthier & Roulin, 2020). The correlation between the two latent variables of the bifactor model was set to zero. For such bifactor models, both latent variables should explain a significant portion of variance (e.g., Lozano 2015; Ren, Gong et al., 2017) adding additional relevance beyond increased model fit.

Figure 2

Illustrates the one factor and bifactor model


Item-position effect / change in toggle rate
.01
.02
...
.25
.26
.27
i_1
i_2
...
i_25
i_26
i_27
i_1
i_2
...
i_25
i_26
i_27
Reasoning ability / basic toggle rate
Reasoning ability / basic toggle rate
Model A
Model B

Note. Model A illustrates a one-factor model with one factor for reasoning ability or basic toggle rate. Model B shows a bifactor model where an item-position effect or the change in toggle rate is included in the analysis respectively.

For the second objective of the current study, we analysed whether an item-position effect emerged in the APM score data. To do so, we fit the same two models to the APM score data as to the toggle rate. In the bifactor model one latent variable depicts reasoning ability and the second the item-position effect (Figure 2, Panel B). The modelled increase for the latent variable of the item-position effect describes the growing relation between observed item responses and the item-position effect. For these two models fit to the APM scores we used the threshold free approach introduced by Schweizer (2013) with probability-based covariance matrices, and factor loadings weighed by the item's standard deviation to account for the difference between binary data distribution and normal distribution (Schweizer et al., 2015). For the reasoning ability latent variable this link was creating by setting the starting value of the factor loading estimation equal to the respective items standard deviation (this works just like pre-multiplication, see Rosseel, 2012). For the latent variable depicting the item-position effect this link was created by multiplying the value for the linear increase with the respective item standard deviation.

For the third objective of the current study, the best measurement model for the APM scores and the toggle rate, was selected to fit a final model. In the final model the relation between reasoning ability and basic toggle rate was examined and their relation with a possible change in toggle rate and / or the item-position effect.

Each measurement model was evaluated with descriptive fit indicines. As a Goodness-of-Fit measure the Comparative Fit Index (CFI) is frequently used, higher value indicating a better fit relative to the independence model. For the CFI a value above 0.90 indicates an acceptable fit, and values above 0.95 a good fit. For the overall model fit the Root Mean Squared Approximation (RMSEA) should be below 0.08 for an acceptable fit and below 0.06 for a good fit. The RMSEA, as its name implies focuses on the error of approximation. The error of approximation is indicative of the lack of fit of the calculated model compared to the model based on the population covariance matrix. The RMSEA is assumed to be relatively independent of sample size and favours parsimonious models, which is of special interest for the current study (Schermelleh-Engel et al., 2003).

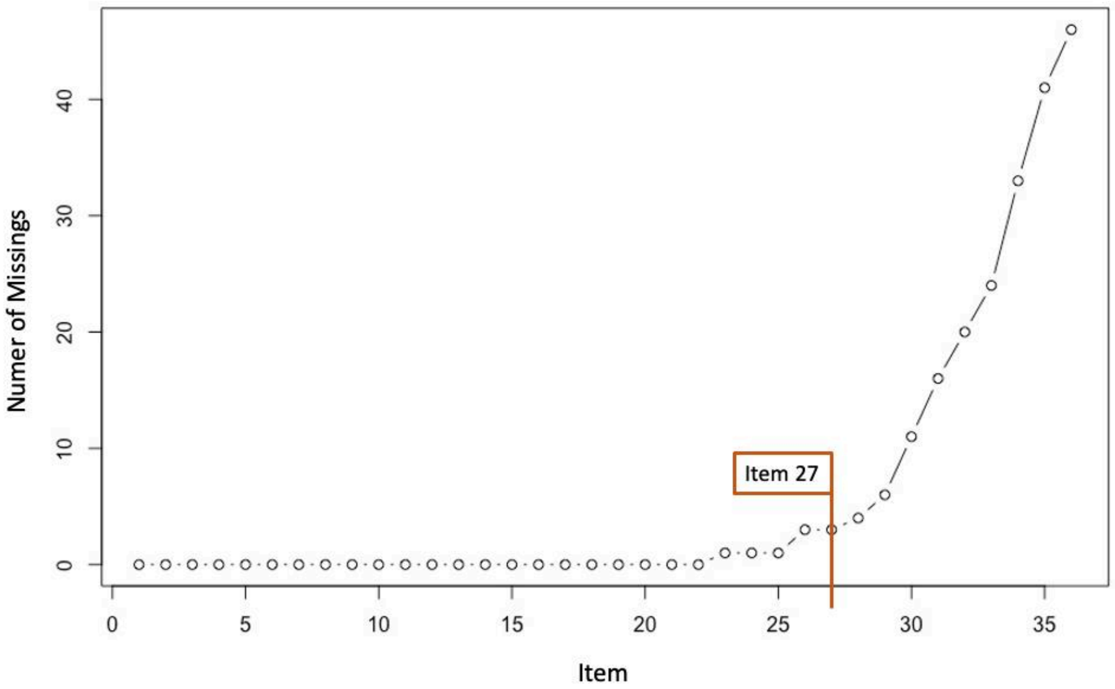
Models were compared with the Akaike Information Criterion (AIC). The AIC adjusts for parsimony, making it an important criterion when comparing competing one-factor and bifactor models. Lower values indicate better fit (Schweizer 2010). In addition to the AIC the Chi-square difference test to compare models. Variances of latent variables were scaled according to Schweizer, Troche, et al. (2019). Data and R-script with the exact factor loadings for the analysis can be accessed on XXX.

Results

Due to technical difficulties only 210 participants had complete data. One participant was excluded from further analysis because of missing information about their age. Six participants had faulty eye tracking data, caused by a recording failure. With the strict time limit implemented in the study, not all participants managed to complete all 36 items. Only 163 participants completed all items, indicating that results toward the end are clearly influenced by the implemented time limit (see Figure 3). Yet all 210 participants completed the first 22 items.

Figure 3

Number of participants that did not solve an item due to implemented time limit



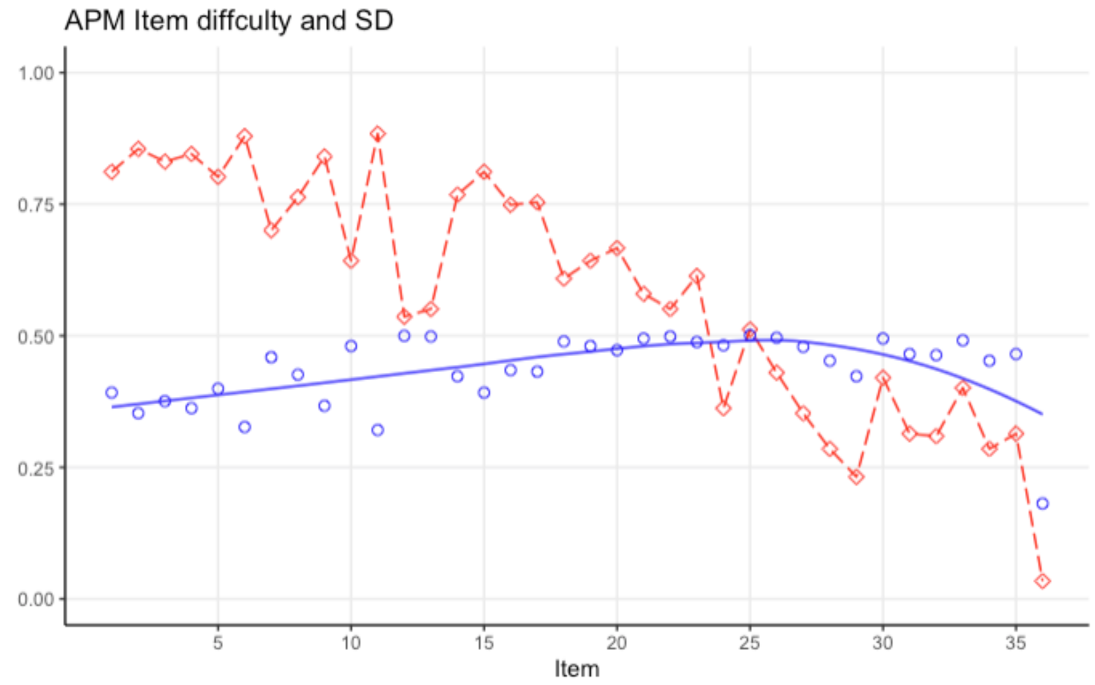
Note. Given are the number of missing values per item due to the implemented time limit.

To only include 22 items in the analysis would increase the risk of not detecting a shift in strategy observed by Gontheir and Roulin (2020). This is also true for the item-position effect, as theory suggests it develops over the course of a test. Deciding on a too late cut off (e.g., item 30), would come at a cost of power if we would want to exclude all participants that did not complete all items used for the analysis (n = 199). Also, a sample with a too late cut off, would include a disproportionate number of participants that worked through the APM items at a faster pace. Therefore, we decided the best trade-off would be to include the first 27 items for the analysis. This included a total of 207 participants in the final analysis (different cut offs reveal the same pattern of results, see supplementary for more details).

The scores in the APM of the remaining 207 participants show that for the selected sample item difficulty increased (lower value in Pi) and the standard deviation also increased throughout the test (see Figure 4). The mean accuracy across all 27 items of the sample was 18.34 with a standard deviation of 5.63. Internal consistency for the APM in the analyzed sample was good (Cronbach’s alpha = 0.87).

Figure 4

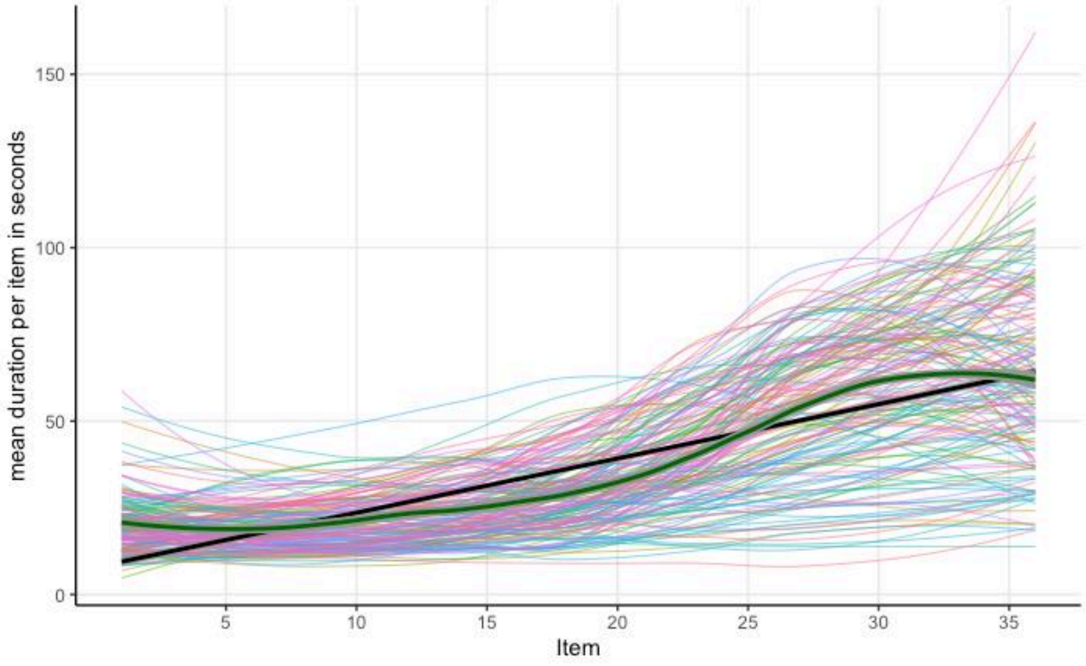
Difficulty (Pi) and standard deviation for all items



Note. Difficulty is depicted by the dashed line and little squares. Little circles and the solid polynomial regression depict standard deviation.

Figure 5

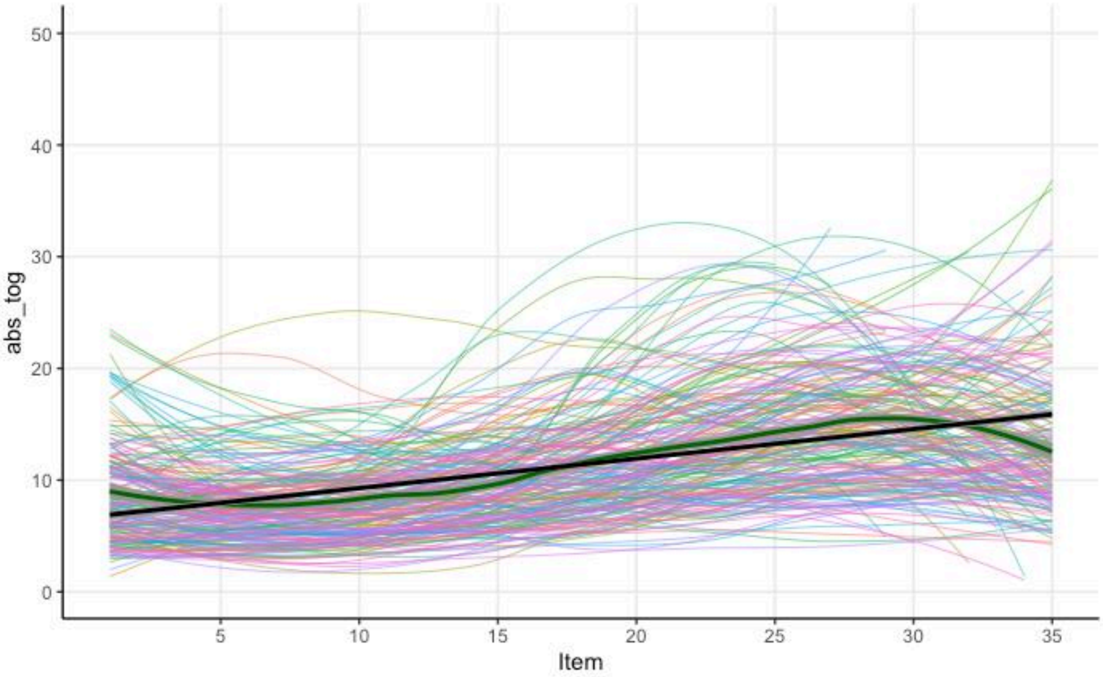
Mean item latency for all items



Note. Linear and local polynomial regression fitted line for analyzed sample and colored for each participant.

Figure 6

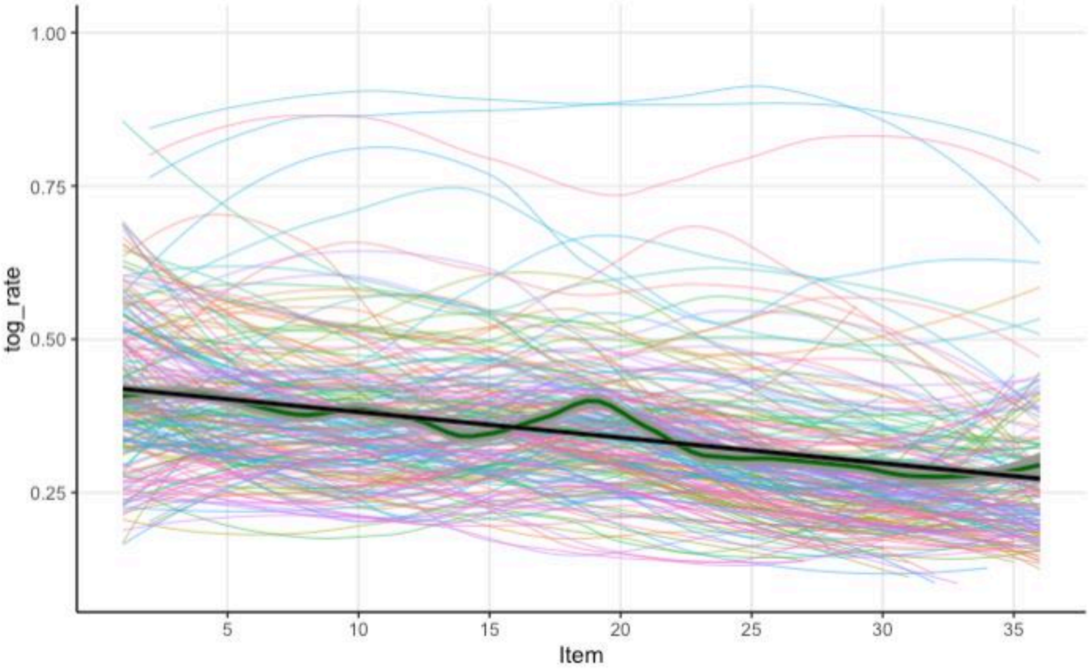
Absolute numbers of Toggles for the 27 analyzed items and participants



Note. Linear (black) and local polynomial (green) regression fitted line for analyzed sample and colored for each participant.

Figure 7

Toggle rate for 27 analyzed items and each participant



Note. Linear (black) and local polynomial (green) regression fitted line for analyzed sample and colored for each participant.

Item latency for each item increased (Figure 5) throughout the test. On average participants had 14.7 minutes to complete all 27 items with a standard deviation of 5.53 minutes, indicating that most participants finished the 27 items clearly below the set time limits of 30 minutes. Within the full sample a linear increase of item latencies throughout the APM can be detected (black line in Figure 5). Fitting a local polynomial regression, we can see that for the first items, latencies seem stable, then increase around approximately item 12, then somewhat flatten out around item 30 and decrease for the last few items.

The absolute number of Toggles divided by the respective item latency equals the analyzed Toggle Rate. The absolute number of Toggle increased from item to item as depicted by the linear regression (in black) in Figure 6. The local polynomial regression (in green) did show a slight drop in absolute numbers of Toggles for the last few items.

The Toggle Rate had high reliability (Cronbach's alpha = 0.95) and interestingly a slight decrease during test completion was observed (see Figure 7). A decrease in Toggle Rate would suggest less response elimination and more constructive matching. We did not observe an increase in standard deviation for the toggle rate.

All the models calculated to determine the best measurement model are summarized in Table 1 and illustrated in Figure 2. The Toggle Rate showed the best fit for the bifactor model where a linear increase for the additional latent variable captures a change in toggle rate (Model B). Fit indices indicate acceptable (CFI) and good fit (RMSEA). Both latent variables explain a significant portion of variance (basic toggle rate: $\varphi = 0.012$, $z = 2.936$, $p < .003$; change in toggle rate: $\varphi = 0.0634$, $z = 4.186$, $p < .001$), and the model shows the lowest AIC.

The bifactor model for the APM scores shows a better fit compared to the one-factor model (smaller AIC). Fit indices for Model B on the APM scores indicate acceptable (CFI) and good fit (RMSEA, SRMR). Both latent variables explain a significant portion of variance (reasoning: $\varphi = 0.0287$, $z = 3.228$, $p = .001$; item-position effect: $\varphi = 0.142$, $z = 3.899$, $p < .001$).

Table 1

	χ^2 *(df)*	*p*	CFI	RMSEA	SRMR	AIC
APM scores						
one-factor model / Model A	413.48 *(324)*	.0006	0.907	0.038	0.059	5628
bifactor model / Model B	394.99 *(323)*	0.004	0.925	0.034	0.059	5610
Toggle rate						
one-factor model / Model A	460.39 *(324)*	<.001	0.928	0.047	0.047	-5548
bifactor model / Model B	430.57 *(323)*	<.001	0.943	0.042	0.047	-5577
Model with toggle rate and APM scores						
Final model	1672.25 *(1371)*	<.001	0.898	0.034	0.060	-75

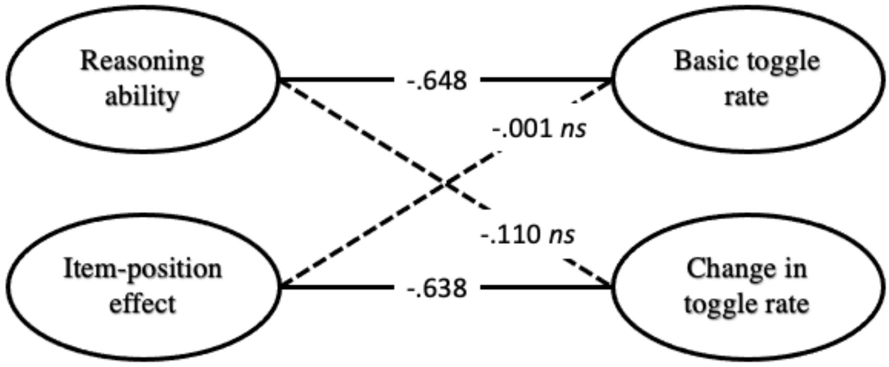
Note. Model A is the respective one-factor model. Model B includes a second latent variable with a linear increase and Modell C with a quadratic increase.

For both the Toggle Rate and the APM scores, the bifactor model improved data description. Even the AIC specifically penalizing less parsimonious models, preferred both bifactor models. Additionally, the added latent variable for the bifactor model described a significant amount of variance in the observed variables in addition to the initially assumed latent variable, further highlighting its relevance. Hence the final model was calculated combining the two best fitting models (Figure 8). The model had acceptable to good fit and all latent variables explained a significant portion of variance. The basic toggle rate: φ = 0.012, $z = 2.967$, $p < .003$; toggle rate variability: φ = 0.034, $z = 4.173$, $p < .001$, reasoning: φ = 0.302, $z = 3.366$, $p = .001$; item-position effect: φ = 0.142, $z = 3.907$, $p < .001$ all explained a significant portion of variance, therefore contributing to describing the data in the model. Exploring modification indices, to possibly increase model fit, 13 modifications with a suggested change of 10 – 18 in χ^2 value were indicated. Several APM scores and Toggle Rates showed correlations among themselves and between each other. The only regression indicated was between the observed score variance of item 6 and the latent variable depicting innate Toggle Rate. Hence, we decided to leave the final model as is.

As can be seen in Figure 8, the latent variable for reasoning ability was strongly correlated with basic toggle rate, but not the change of toggle rate. The item-position effect was strongly correlated with the change in toggle rate, but not the basic toggle rate. To illustrate and better understand these relations, descriptive graphics were created. For the graphs, the factor scores of the final model were extracted. Then, for the participants with the 50 highest and the 50 lowest factor scores the toggle rate throughout the test was illustrated separately for the respective latent variables.

Figure CC

Correlations between latent variables of the final model



We can see in Figure 9 in the top left corner the toggle rate for participants with high factor scores on the latent variable of reasoning ability (red line) and participants with low factor scores on reasoning ability (blue line). Participants with higher factor scores on the reasoning ability variable show a lower toggle rate for each item compared to participants with low reasoning ability factor scores. This means participants rather engaged in response elimination when their reasoning ability was low. This is in line with the conclusion of Vigneau et al., (2006)

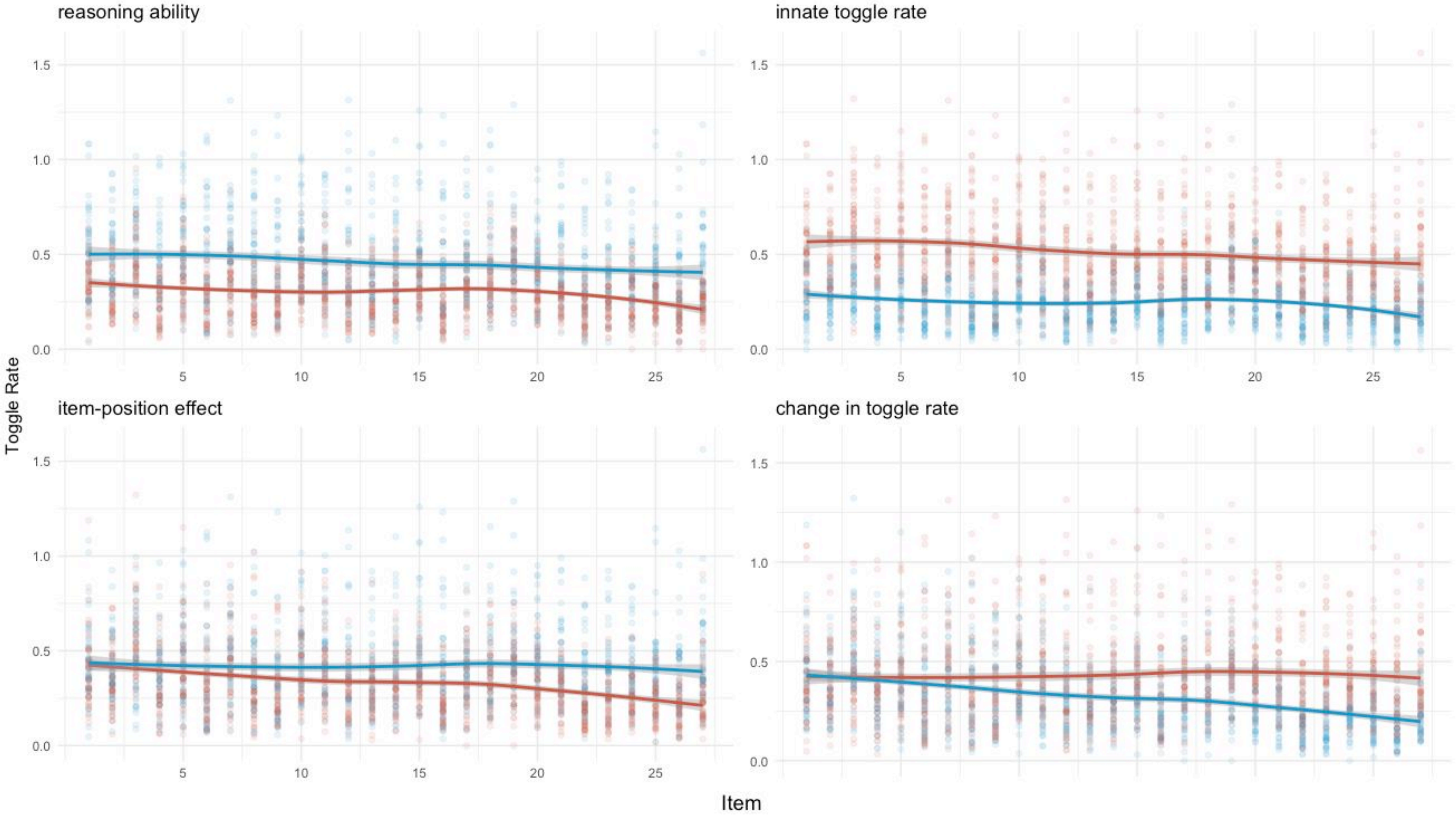
If we look at the latent variable of the basic toggle rate (top right), participants with a higher factor score (red line) also showed an overall higher toggle rate. This indicates that the latent variable for the basic toggle rate reflects general toggle rate and individuals with a high toggle rate, that most likely engage in response elimination have higher factor scores compared to individuals that rather engage in constructive matching.

Participants with low factor scores (blue line) on the item-position effect (bottom left), show no change in their overall toggle rate, meaning there is most likely no change in strategy throughout the APM. Participants with high factor scores (red line), reduce their toggle rate throughout the APM, reflecting a change in strategy. This indicates that participants with high factor scores on the latent variable depicting the item-position effect, adapt their strategy during test taking, and shift to making fewer toggles, that is, less response elimination or more constructive matching.

Participants with high factors scores (red line) on change in toggle rate (bottom right) show no explicit change in their overall toggle rate. Participants with low values on the other hand, show a decrease in their overall toggle rate.

Figure 9

Toggle rate for participants with high or low values on the latent variables



Note. Blue lines represent participants with low factor scores on the latent variable. Red lines represent participants with high factor scores on the depicted latent variable.

These results are also reflected in the correlations of the final model, if less descriptive. A low value in change of toggle rate is related to a high value on the item-position effect. And a low value in reasoning ability is related to a high value in innate Toggle Rate. Hence the large negative correlations between these latent variables.

Discussion

Results of the current study provide new information about strategy use during the completion of the APM. The first objective was to assess whether individual differences in strategy, operationalized trough toggle rate, can be identified and modelled with a confirmatory factor analysis using a fixed-links approach. Data description indeed improved, after adding a second latent variable to the model. This indicates that accounting for changes in toggle rate in addition to innate differences in toggle rate further explain a unique amount of variance in the data. It is therefore fair to assume, that there is as Vigneau et al., (2006) described a general difference between individuals as to what strategy they mostly use, but also as Gonthier and Roulin (2020) found (albeit with a somewhat different trajectory), a shift or change occurring of said strategy use during test completion.

The second objective simply aimed to examine whether an item-position effect indeed was present in the APM score data, since this seemed to be a common occurrence within reasoning ability measures such as the APM (e.g., Schweizer et al., 2012; Zeller et al., 2017). Present results do suggest the presence of an item-position effect in the APM score data. An improved model fit was found for the bifactor model and importantly both latent variables, reasoning ability and the item-position effect explained a significant amount of variance in the data.

The third objective of the current study was to analyze whether the toggle rate was related to reasoning ability, and if a change in toggle rate throughout test completion occurred, whether this change is related to reasoning ability and / or the item-position effect. The final model to investigate this objective showed a strong negative correlation between reasoning ability and innate toggle rate underlining the findings of Vigneau et al., (2006). Our data supports that higher reasoning ability scores coincide with a lower toggle rate. Meaning, with higher reasoning ability more constructive matching is used.

Additionally, the final model showed a similarly large negative correlation between the item-position effect and the change in toggle rate. This indicates that individuals with a pronounced item-position effect, reduce their toggle rate through the test (have a larger negative change – since less pronounced item-position effect seems to elicit no change in toggle rate), changing their strategy use (at least to some degree). Taking a closer look at common reasoning ability measures the strong relation between the item-position effect and the change strategy use becomes inherently logical.

Most reasoning tests rely on a limited set of different rules, that are combined to create different items (e.g., Carpenter et al., 1990). It is therefore safe to assume that when the same rule is used repeatedly in a test, some sort of learning takes place. Several studies support this assumption (Bui & Birney, 2014; Carlstedt et al., 2000; Verguts & De Boeck, 2000). In a follow-up study Verguts and De Boeck (2002) even observed that the learning effects seem to be rule specific.

Regarding the item-position effect studies support the premise that continuous learning underlies the item-position effect. Results of Ren et al. (2014) and Schweizer et al. (2021) showed modest to strong correlational relationships between the item-position effect and the performance on complex learning tasks. Therefore, individuals with a pronounced item-position effect most likely are effective at rule learning during test completion.

To learn rules during test taking, constructive matching seems the most expedient strategy. When individuals spend time on the problem matrix, they engage in a systematic analysis of the stimulus (Snow, 1978) and therefore are able analyze the rules necessary to solve the item. Subsequently they then can create an answer in their mind. Hence, individuals are more mentally engaged with the rules when they use constructive matching rather than response elimination.

Constructive matching fosters rule learning whereas response elimination does not. Constructive matching should facilitate solving subsequent items. This leads us to conclude, that individuals who engage in constructive matching and learn the underlying rules of items during test taking, are more likely to continue applying constructive matching throughout the test. Proceeding from this line of thought, individuals engaging in constructive matching can be expected to show a more pronounced item-position. This is reflected in the strong negative correlation between the item-position effect and the change in strategy use.

Pointing to the established positive relation of constructive matching and the performance on the APM (e.g., Jastrzebski et al., 2018), it is important to note, that with simple training Hayes et al., (2015) found that in a test – retest setting a third of the variance in score gains was due to strategy use. Also, when participants were given the rules necessary to solve the items beforehand, rendering the learning of rules mute, participants used more constructive matching during test completion (Loesche et al., 2015). Gonthier and Thomassin

(2015) found that the use of the more expedient strategy, constructive matching, can be successfully manipulated, albeit Mitchum and Kelley (2010) did not find enhanced performance through strategy training in their data.

Research therefore suggests that under certain circumstances individuals can be supported in selecting the more successful strategy of constructive matching. This information is prudent, since differences in strategy use are already observed in young children (Starr et al., 2018). The common use of reasoning ability measures to determine what level of education seems suitable for young adults (e.g., Sonnleitner et al., 2013; Gomez-Veiga et al., 2018), or what outlook a person has in future job performance (e.g., Salgado et al., 2003), underlines the necessity for this line of research.

It seems that simple training in strategy use could enhance performance in reasoning ability measures which in turn could impact one's future. Of course, additional studies on the topic are necessary to determine whether score gains through strategy use training would be substantial enough to impact live outcomes. Nevertheless, the thought is enticing.

Strategy and Toggle Rate

For further understanding of the results, a closer inspection of the toggle rate in general is warranted. Overall, a decrease in toggle rate was observed, indicating that participants used more constructive matching towards the end of the APM. This is the opposite of what Snow (1980), or Bethel-Fox et al. (1984) concluded from their results. They reported a decrease of constructive matching on difficult items. With the nature of the APM later items are more difficult than earlier items. Therefore, we would have expected an increase in toggle rate, created by more response elimination and less constructive matching.

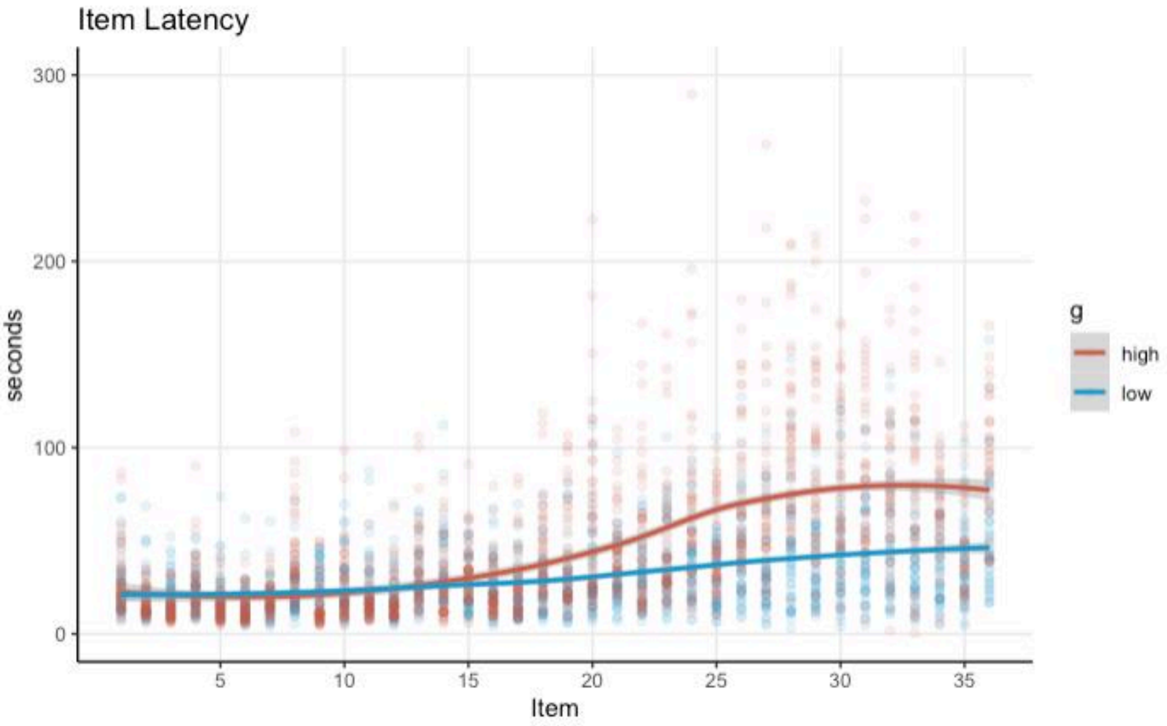
Such an increase in response elimination was found by Gonthier and Roulin (2020). In their study individuals either solved the odd or even items of the APM in progressive order and answered two questions about their strategy after each item. With 100 participants for each version (odd/even) of the APM administered, results are supported by enough statistical power and are reliable. Nevertheless, implementing the same measure for reasoning ability, having a similar overall sample size, different results emerge. The only difference in administration was the set time limit and the means of measuring strategy (eye-tracking vs questionnaire).

The set time limit to complete the APM, requires a closer look at item latencies. Item latencies in Gonthier and Roulin (2020) increased during APM completion upon item 30, and then decreased. Authors believe this to be a sign of disengagement of participants as items became too taxing. This pattern was not replicated within our data. Item latencies for the full sample also increased upon item 30, but then seem to flatten out (see Figure 5). Of course, it must be mentioned, that with the time limit of 30 minutes, not all participants have worked on the later items. Therefore, item latencies for the last items are somewhat biased, since only individuals are included that worked through the items with a faster pace. Nonetheless, this points towards a qualitative difference between samples, which could be a source for the differences in results.

A further visual inspection of item latencies supports the idea of qualitative differences between presented results. Our data shows a different change in item latencies throughout test taking between high and low reasoning ability individuals (see Figure 10). While all participants show an increase in item latencies, these increases start to deviate from each other around item 15. Participants with high reasoning ability factor scores show a larger increase in item latency and this difference between participants with high and low reasoning ability factor scores grows, up until approximately item 30 and then seems to shrink for the last few items. This makes sense, since as displayed in Figure 9, participants with higher reasoning ability progressively used more constructive matching, which requires a more detailed analysis of the problem matrix, possibly requiring more time to solve more difficult items. Nevertheless, further investigations as to what circumstances led to the difference between the present results and findings of Gonthier and Roulin (2020) are paramount.

Figure 10

Item Latencies for participants with high or low reasoning ability factor scores



Note. Given are item latencies in seconds for the participants with the 50 highest (in red) and 50 lowest (in blue) reasoning ability factor scores of the final model.

Our results also diverge from other work. Consulting work based on eye-tracking data, Vigneau et al. (2006) found a positive relation between strategy use and reasoning ability but not any shift or change in strategy use. This does not coincide with our observations. Possibly the study design, using a subset of APM items without easy items, and a sample size of 55 participants did not allow for the detection of an adaption in strategy. With their sophisticated analysis of eye tracking data, authors concluded, that an inherit difference in reasoning ability influences the outcome of what strategy an individual engages in. When consulting the results from the full model of the current study, we can see in Figure 9 (top left), that if we differentiate between basic toggle rate and change in toggle rate, that indeed individuals with low reasoning ability have a higher toggle rate compared to high ability individuals. These results are therefor in line with the conclusions of Vigneau et al. (2006). But in addition, we observed individual differences of change in toggle rate.

Some individuals did not change their strategy throughout the test, while others showed a decrease in toggle rate, indicating more constructive matching. Interestingly, this change in toggle rate was not associated with reasoning ability when the item-position effect was accounted for (Figure 8). If we disregard the existence of the item-position effect, a small correlation (*r* =-.28) between change in toggle rate and reasoning ability can be detected.

This would be in line with the findings of Liu et al. (2023) where an increase in constructive matching was found for high ability group, whereas no change for medium and slight decrease for low ability groups occurred. Their regression analysis also revealed that the main effect of item-difficulty was not significant in predicting the usage of constructive matching. The interaction term of intelligence and item-difficulty on the other hand was significant, indicating that ability effects the relation between constructive matching and item-difficulty.

Unfortunately, item-difficulty and item-position are strongly confounded in the APM, making a clear separation of the two somewhat impossible. Alas, the current results cannot provide further evidence along this line of research. Also, the direct comparison of the present results with the conclusions from Liu et al. (2023), albeit interesting are to be overinterpreted, since our analysis clearly showed an item-positing effect in the data and theirs did not. Therefore, present data does not support the assumption that, reasoning ability coincides with a change in strategy (i.e., toggle rate).

Current results show that the change in strategy is strongly related to the item-position effect. Individuals with a high manifestation of the item-position effect show low values in the latent variable depicting change in toggle rate. The analysis of Vigneau et al. (2006) or Liu et al. (2023) did not include an item-position effect, which could be a reason for the diverging results.

As we can see in Figure 9 low values on said latent variable indicate a change to more constructive matching throughout test completion. Therefore, individuals with a strong influence of item-position start using more constructive matching while solving the APM. These results are in line with the learning hypothesis brought forward by Ren et al. (2014) and support the presented conclusion, that constructive matching fosters rule learning.

Looking outside the realm of eye-tracking studies regarding strategy more recent work has identified a third strategy. Namely the isolate-eliminate strategy was identified by Jarosz et al. (2019) using think out loud protocols. Here individuals isolate a group of response alternatives based on one characteristic that is easy to identify by the individual to be a wrong answer. Basically, bad lures, that are easy to discern are isolated in groups and eliminated, leaving fewer and fewer options to choose from, increasing the likelihood for a correct answer.

Li et al. (2022) also found a third strategy running a latent profile analysis on in their questionnaire data. Interestingly the questionnaire aimed to identify to what extent an individual uses constructive matching or response elimination. Questions were translated from the set used by Jasterzebski et al., (2018). The additionally identified strategy had individuals scoring high on questions regarding constructive matching as well as response elimination. While the authors found differences in APM performance for the three groups, there was no difference regarding the toggle rate between the response elimination and the third group.

It seems that toggle rate does not distinguish between the response elimination strategy and the possible third strategy, isolate-eliminate. Also, from the current state of research it is not yet clear, whether the third strategy found by Jarosz et al., (2019) within their think out loud protocols is the same additional strategy found by Li et al., (2022). Further research, maybe of a more explorative nature regarding strategy might shed light on this conundrum.

References

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*(3), 205–238. https://doi.org/10.1016/0160-2896(84)90009-6

Bojko, A. (*2013*). Eye Tracking, the User Experience: A Practical Guide to Research. Brookling, New York: Rosenfield Media.

Bui, M., & Birney, D. P. (2014). Learning and individual differences in Gf processes and Raven's. *Learning and Individual Differences*, *32*, 104–113. https://doi.org/10.1016/j.lindif.2014.03.008

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Carlstedt, B., Gustafsson, J. E., & Ullstadius, E. (2000). Item sequencing effects on the measurement of fluid intelligence. *Intelligence*, *28*(2), 145–160. https://doi.org/10.1016/S0160-2896(00)00034-9

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404. https://doi.org/10.1037/0033-295X.97.3.404

Gómez-Veiga I, Vila Chaves JO, Duque G and García Madruga JA (2018) A New Look to a Classic Issue: Reasoning and Academic Achievement at Secondary School. *Front. Psychol*. 9:400. https://doi.org/10.3389/fpsyg.2018.00400

Gonthier, C., & Roulin, J.-L. (2020). Intraindividual strategy shifts in Raven’s matrices, and their dependence on working memory capacity and need for cognition. *Journal of Experimental Psychology: General*, *149*(3), 564–579. https://doi.org/10.1037/xge0000660

Gonthier, C., & Thomassin, N. (2015). Strategy use fully mediates the relationship between working memory capacity and performance on Raven's matrices. *Journal of Experimental Psychology: General*, *144*(5), 916–924. https://doi.org/10.1037/xge0000101

Gottfredson, L. S., & Deary, I. J. (2004). Intelligence Predicts Health and Longevity, but Why? Current Directions in Psychological Science, 13(1), 1–4. https://doi.org/10.1111/j.0963-7214.2004.01301001.x

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

Jarosz, A. F., Raden, M. J., & Wiley, J. (2019). Working memory capacity and strategy use on the RAPM. Intelligence, 77, 101387. https://doi.org/10.1016/j.intell.2019.101387

Jastrzębski, J., Ciechanowska, I., & Chuderski, A. (2018). The strong link between fluid intelligence and working memory cannot be explained away by strategy use. *Intelligence*, *66*, 44–53. https://doi.org/10.1016/j.intell.2017.11.002

Laurence, P. G., Mecca, T. P., Serpa, A., Martin, R., & Macedo, E. C. (2018). Eye Movements and Cognitive Strategy in a Fluid Intelligence Test: Item Type Analysis. *Frontiers in Psychology*, *9*, 380. https://doi.org/10.3389/fpsyg.2018.00380

Li, C., Ren, X., Schweizer, K., & Wang, T. (2022). Strategy use moderates the relation between working memory capacity and fluid intelligence: A combined approach. *Intelligence*, *91*, 101627. https://doi.org/10.1016/j.intell.2022.101627

Liu, Y., Zhan, P., Fu, Y., Chen, Q., Man, K., & Luo, Y. (2023). Using a multi-strategy eye-tracking psychometric model to measure intelligence and identify cognitive strategy in Raven's advanced progressive matrices. *Intelligence*, *100*, 101782. https://doi.org/10.1016/j.intell.2023.101782

Loesche, P., Wiley, J., & Hasselhorn, M. (2015). How knowing the rules affects solving the Raven Advanced Progressive Matrices Test. *Intelligence*, *48*, 58–75. https://doi.org/10.1016/j.intell.2014.10.004

Lozano, J. H. (2015). Are impulsivity and intelligence truly related constructs? Evidence based on the fixed-links model. *Personality and Individual Differences*, *85*, 192–198. https://doi.org/10.1016/j.paid.2015.04.049

Martarelli, C. S., & Mast, F. W. (2013). Eye movements during long-term pictorial recall. *Psychological research*, *77*, 303–309. https://doi.org/10.1007/s00426-012-0439-7

Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: Constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 699–710. https://doi.org/10.1037/a0019182

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, *51*, 195–203. https://doi.org/10.3758/s13428-018-01193-y

Raden, M. J., & Jarosz, A. F. (2022). Strategy Transfer on Fluid Reasoning Tasks. *Intelligence*, *91*, 101618. https://doi.org/10.1016/j.intell.2021.101618

Raven, J. C., Raven, J., & Court, J. H. (1998). *Advanced Progressive Matrices [Measurement Instrument]*. https://www.testzentrale.ch/shop/advanced-progressive-matrices.html

Ren, X., Gong, Q., Chu, P., & Wang, T. (2017). Impulsivity is not related to the ability and position components of intelligence: A comment on Lozano (2015). *Personality and Individual Differences*, *104*, 533–537. https://doi.org/10.1016/j.paid.2016.09.007

Ren, X., Schweizer, K., Wang, T., Chu, P., & Gong, Q. (2017). On the relationship between executive functions of working memory and components derived from fluid intelligence measures. *Acta Psychologica*, *180*, 79–87. https://doi.org/10.1016/j.actpsy.2017.09.002

Ren, X., Schweizer, K., Wang, T., & Xu, F. (2015). The prediction of students' academic performance with fluid intelligence in giving special consideration to the contribution of learning. *Advances in Cognitive Psychology*, *11*(3), 97–105. https://doi.org/10.5709/acp-0175-z

Ren, X., Wang, T., Altmeyer, M., & Schweizer, K. (2014). A learning-based account of fluid intelligence from the perspective of the position effect. *Learning and Individual Differences*, *31*, 30–35. https://doi.org/10.1016/j.lindif.2014.01.002

Revelle, W., & Revelle, M. W. (2015). Package 'psych'. *The comprehensive R archive network*, *337*(338), 161–165.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1–36.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., De Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of applied psychology*, *88*(6), 1068.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online*, *8*(2), 23-74.

Schweizer, K. (2006). The fixed-links model for investigating the effects of general and specific processes on intelligence. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 2*(4), 149–160. https://doi.org/10.1027/1614-2241.2.4.149

Schweizer, K. (2010). Some guidelines concerning the modeling of traits and abilities in test construction. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000001

Schweizer, K. (2013). A threshold-free approach to the study of the structure of binary data. *International Journal of Statistics and Probability*, *2*(2), 67. https//doi.org/10.5539/ijsp.v2n2p67

Schweizer, K., Reiss, S., Schreiner, M., & Altmeyer, M. (2012). Validity improvement in two reasoning measures following the elimination of the position effect. *Journal of Individual Differences*. *33*(1), 54–61. https://doi.org/10.1027/1614-0001/a000062

Schweizer, K., Ren, X., Wang, T. (2015). A Comparison of Confirmatory Factor Analysis of Binary Data on the Basis of Tetrachoric Correlations and of Probability-Based Covariances: A Simulation Study. In: Millsap, R., Bolt, D., van der Ark, L., Wang, WC. (eds) Quantitative Psychology Research. Springer Proceedings in Mathematics & Statistics, vol 89. Springer, Cham. https://doi.org/10.1007/978-3-319-07503-7_17

Schweizer, K., & Troche, S. (2018). Is the factor observed in investigations on the item-position effect actually the difficulty factor? *Educational and Psychological Measurement*, *78*(1), 46–69. https://doi.org/10.1177/0013164416670711

Schweizer, K., & Troche, S. (2019). The EV Scaling Method for Variances of Latent Variables. *Methodology*, *15*(4), 175–184. https://doi.org/10.1027/1614-2241/a000179

Schweizer, K., Troche, S., Rammsayer, T., & Zeller, F. (2021). Inductive reasoning and its underlying structure: Support for difficulty and item position effects. *Advances in Cognitive Psychology*, *17*(4), 274–283. https://doi.org/10.5709/acp-0336-5

Snow, R. E. (1978). Eye Fixation and Strategy Analyses of Individual Differences in Cognitive Aptitudes. In A. M. Lesgold, J. W. Pellegrino, S. D. Fokkema, & R. Glaser

(Eds.), *Cognitive Psychology and Instruction* (pp. 299–308). Springer US. https://doi.org/10.1007/978-1-4684-2535-2_27

Snow, R. E. (1980). Aptitude Processes. In R. E. Snow, P.-A. Federico, & W. E. Montague (Eds.), *Aptitude, learning, and instruction: Cognitive process analyses of aptitude* (Vol. 1, pp. 27–64). Erlbaum.

Sonnleitner, P., Keller, U., Martin, R., & Brunner, M. (2013). Students' complex problem-solving abilities: Their structure and relations to reasoning ability and educational success. *Intelligence*, *41*(5), 289–305. https://doi.org/10.1016/j.intell.2013.05.002

SR Research. (2016). Eyelink 1000 plus [Apparatus and software]. https://www.sr-research.com/eyelink1000plus.html

Starr, A., Vendetti, M. S., & Bunge, S. A. (2018). Eye movements provide insight into individual differences in children's analogical reasoning strategies. *Acta Psychologica*, *186*, 18–26. https://doi.org/10.1016/j.actpsy.2018.04.002

Troche, S. J., Wagner, F. L., Schweizer, K., & Rammsayer, T. H. (2016). The Structural Validity of the Culture Fair Test Under Consideration of the Item-Position Effect. *European Journal of Psychological Assessment*, *35*(2), 182–189. https://doi.org/10.1027/1015-5759/a000384

Verguts, T., & De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement*, *24*(2), 151–162. https://doi.org/10.1177/01466210022031589

Verguts, T., & De Boeck, P. (2002). The induction of solution rules in Raven’s Progressive Matrices Test. *European Journal of Cognitive Psychology*, *14*(4), 521–547. https://doi.org/10.1080/09541440143000230

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, *34*(3), 261–272. https://doi.org/10.1016/j.intell.2005.11.003

von Gugelberg, H. M., Schweizer, K., & Troche, S. J. (2021). The dual mechanisms of cognitive control and their relation to reasoning and the item-position effect. *Acta Psychologica*, *221*, 103448. https://doi.org/10.1016/j.actpsy.2021.103448

Zeller, F., Reiss, S., & Schweizer, K. (2017). Is the item-position effect in achievement measures induced by increasing item difficulty? *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 745–754. https://doi.org/10.1080/10705511.2017.1306706

Appendix C

Study 3: Rule Disruption and the Learning Hypothesis

Experimental evidence for rule learning as the underlying source of the item-position effect in reasoning ability measures

This article is published as:

von Gugelberg, H. M., Schweizer, K., & Troche, S. J. (2025). Experimental evidence for rule learning as the underlying source of the item-position effect in reasoning ability measures. *Learning and Individual Differences. 118,* 102622.

https://doi.org/10.1016/j.lindif.2024.102622

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

Experimental evidence for rule learning as the underlying source of the item-position effect in reasoning ability measures

Helene M. von Gugelberg [a,*], Karl Schweizer [b], Stefan J. Troche [a]

[a] Department of Psychology, University of Bern, Switzerland
[b] Department of Psychology and Sports Sciences, Goethe University Frankfurt, Frankfurt a. M., Germany

ARTICLE INFO

Keywords:
Reasoning ability
Reasoning test
Item-position effect
Rule learning
Confirmatory factor analyses

ABSTRACT

For adequate description of reasoning test data, the consideration of the item-position effect (IPE) as a second latent variable in addition to reasoning ability is often required. The present study investigated the assumption that the learning of rules underlies the IPE. The factorial structure of two figural analogies tests was compared. 429 participants (age: 18–56 years) were randomly assigned to two conditions. In the continuous rule condition, the same rule had to be applied to all items and a typical IPE emerged. In the discontinuous rule condition, rules suddenly changed for the last items. This change led to the disruption of the IPE. A third latent variable was required to describe variance in the last items. Thus, the repetition of rules seems to be a precondition for a continuous IPE across test items. This is first evidence beyond correlations that individual differences in rule learning underlie the IPE.

Educational relevance: Reasoning tests are frequently used as an indicator of (general) intelligence and are valuable predictors of academic achievement. Our online experiment provides evidence for the notion that individual differences in reasoning tests are influenced by ad hoc rule learning during test taking that can be described as a latent variable and separated from the innate reasoning ability. These findings highlight the importance of not only looking at the overall performance but paying more attention to the dynamics of the test taking process itself.

1. Introduction

In structural models of intelligence (e.g., Carroll, 1993; McGrew, 2009), reasoning abilities (at a first stratum) entail the abilities to draw inferences from predetermined premises, to detect relations, to identify rules in a set of figures and to solve logical puzzles or abstract problems (Carroll, 1993). These abilities form the components of fluid intelligence at a second, broader and more abstract stratum defined as the ability to solve novel problems by controlled cognitive processing (McGrew, 2009). This broad ability is closely related to general intelligence at the third (and highest) stratum of hierarchical intelligence models (Johnson, te Nijenhuis, & Bouchard Jr, 2008; Kan, Kievit, Dolan, & van der Maas, 2011; Kvist & Gustafsson, 2008). This close relation between fluid and general intelligence might be the reason why reasoning tests are not only an integral element of the most established test batteries such as the Wechsler tests (Wechsler, 2017) but also the most used instruments when intelligence is measured by a single scale instead of a test battery (Roth & Herzberg, 2008).

Although most reasoning tests use a series of very similar items, they have frequently found to be less homogenous than one might expect. Often a one-factor solution failed to describe the responses to items of a reasoning test well (e.g., Dillon, Pohlmann, & Lohman, 1981; Van der Ven & Ellis, 2000; Vigneau & Bors, 2008). Several reasons for this lack of homogeneity have been proposed primarily for Raven's Advanced Progressive Matrices (RAPM; Raven & Raven, 2003), which are used frequently to measure reasoning in psychological research. Like many measures of reasoning ability, the RAPM consist of a series of similar problems/items.

Possible reasons for the lack of homogeneity of the RAPM were described by Carpenter, Just, and Shell (1990). They highlighted that the number and type of rules differ between items. DeShon, Chan, and Weissbein (1995) emphasized that some RAPM items required visuo-spatial and other items verbal-analytic processes. Embretson (1995) found that modeling two abilities underlying the responses of RAPM items described her data well. One of these abilities was required by all items in a similar way, and the other ability was more relevant for later

* Corresponding author.
E-mail address: helene.vongugelberg@unibe.ch (H.M. von Gugelberg).

https://doi.org/10.1016/j.lindif.2024.102622
Received 8 January 2024; Received in revised form 16 October 2024; Accepted 19 December 2024

than earlier items.

Embretson's (1995) idea of two latent variables underlying the performance on reasoning test items has been revived in the last years by applying confirmatory factor analysis (CFA; e.g., Schweizer, Reiss, Schreiner, & Altmeyer, 2012). With this approach, a bifactor model is fit to the data, where a first latent variable representing reasoning ability is complemented by a second latent variable capturing the increasing (true) item variance from the first to the last item of a reasoning test. In this bifactor model, the second latent variable can be extracted from the same set of items as the first latent variable because its factor loadings are fixed to monotonically increase from the first to the last item (e.g., Zeller, Krampen, Reiss, & Schweizer, 2017). Since the factor loadings of the second latent variable increase with the items' position, this latent variable is referred to as item-position effect (Schweizer, Schreiner, & Gold, 2009). This approach focuses on the systematic variation of data. Different sources of response behavior lead to different patterns of systematic variation that need to be captured by different latent variables reflecting interindividual differences.[1]

A common finding of such studies is that both reasoning ability and the item-position effect explain substantial and unique portions of individual differences in test performance and their concurrent consideration does not only improve the model/data fit but leads to an improved data description (Schweizer et al., 2012; Troche, Wagner, Schweizer, & Rammsayer, 2016). This has been shown for the RAPM (Ren, Schweizer, Wang, Chu, & Gong, 2017) but also for Horn's (1983) sequential reasoning test (Ren, Gong, Chu, & Wang, 2017), Cattell's Culture Fair Test (Troche et al., 2016) and Formann, Piswanger, and Waldherr's (2011) Vienna Matrices Test (von Gugelberg, Schweizer, & Troche, 2021).

The relevance of the item-position effect can not only be seen in its contribution to better data description but also in its predictive value for performance in everyday life. Ren, Schweizer, Wang, and Xu (2015) demonstrated that the item-position effect served as a better predictor for verbal and math grades for secondary school students compared to the latent variable representing reasoning ability. These results suggest that the item-position effect adds to the predictive validity of reasoning tests for school grades above and beyond reasoning ability. It does not simply represent a method effect but individual differences that are psychologically meaningful. Unfortunately, it is still unclear what ability is reflected in the item-position effect.

The items of reasoning tests are often arranged with increasing difficulty. This led to the assumption that the item-position effect reflects item difficulty. This explanation was ruled out by simulation (Schweizer & Troche, 2018) as well as empirical studies (Zeller, Reiss, & Schweizer, 2017). Zeller, Reiss, and Schweizer (2017) presented items of the RAPM in random order. With this random order, item difficulty was independent of item position and the item-position effect still occurred. Additionally, Schweizer, Troche, Rammsayer, and Zeller (2021) found that the difficulty effect was strongly related to the reasoning latent variable while the item-position effect could easily be dissociated from it. These results show that the item-position effect does not reflect increasing item difficulty in reasoning tests.

A more sensible explanation as to why the item-position effect occurs could be that most reasoning tests use a limited number of rules and their variations and combinations (e.g., Carpenter et al., 1990). It is plausible to assume that when a rule is used repeatedly a learning process takes place during test completion. This premise is consistent with for example, Carlstedt, Gustafsson, and Ullstadius (2000). In their study, items of three reasoning tests were presented in two different conditions. In one condition, items of the three tests were sorted according to the test they belonged to. In the other condition, items of the three tests were mixed. In the sorted condition, the later items were more frequently solved correctly compared to the mixed condition. Carlstedt et al. (2000) explained these differences through learning effects from solving earlier items transferring to later items.

Also, Verguts and De Boeck (2000) observed learning effects during the completion of a reasoning test and individual differences therein. In a subsequent study, the same authors found the learning effects in the RAPM to be rule specific (Verguts & De Boeck, 2002). Similarly, Birney, Beckmann, Beckmann, and Double (2017) identified learning trajectories in the RAPM, that were associated with item position but not with item difficulty.

In two previous studies, the RAPM items were divided in two distinct sets (Harrison, Shipstead, & Engle, 2015; Wiley, Jarosz, Cushen, & Colflesh, 2011). One set consisted of items, where the underlying rules were used for the first time (new-rule items), while the other set contained items, where the rules were not new but used a second time (repeated-rule items). In both studies, the mean scores indicated that the repeated-rule items were easier compared to new-rule items (this was not tested for significance) possibly pointing to a learning effect. Harrison et al. (2015) additionally found that the repeated-rule items correlated stronger with working memory capacity than the new-rule items and interpreted their results in terms of learning efficiency.

Proceeding from the evidence for learning during test completion enabled by rule repetition, Ren, Wang, Altmeyer, and Schweizer (2014) put forward the learning hypothesis for the item-position effect. This hypothesis states that the learning of rules from item to item leads to increasing individual differences across a series of similar items. This increase is then reflected in the latent variable representing the item-position effect. Ren et al. (2014) also provided empirical support for the learning hypothesis. They found a high correlational relationship between the item-position effect in the RAPM and the performance in a complex-rule learning task. This correlation was conceptually replicated by Schweizer et al. (2021).

However, correlational relationships cannot be interpreted in a causal manner. The detected correlations between the item-position effect and complex learning could also be driven by other common cognitive processes involved in both the item-position effect and complex rule learning. Working memory updating, for example, has been shown to be related to the item-position effect (Ren, Schweizer, et al., 2017) as well as learning ability (Gijselaers, Meijs, Neroni, Kirschner, & de Groot, 2017; Ropovik, 2014). Therefore, from the observed correlation between the item-position effect and complex rule learning (Ren et al., 2014; Schweizer et al., 2021) it cannot be concluded that it is indeed learning that creates this connection. From this point of view, the learning hypothesis is theoretically reasonable, but there is no evidence beyond such of correlational analyses to support it.

The aim of the current study was to provide more direct evidence for the learning hypothesis by using an experimental design with two conditions. For the *continuous rule condition*, we created a set of 24 figural analogies where the same rule was applied throughout the whole test. Since the rule could be applied to different elements, item difficulty increased by applying the rule to multiple elements in a single item. If the learning of rules during test completion underlies the item-position effect, the item-position effect should emerge across the items of this set of figural analogies.

In a second condition, the *discontinuous rule condition,* the first 18 items were identical to the continuous rule condition and should also elicit an item-position effect. However, for the last six items, new rules had to be applied to solve the items correctly. This introduction of new rules, rules which could not have been learned through earlier items, should disrupt the item-position effect, if it is truly caused by the learning of rules during test taking. We assume another item-position effect to emerge from the last six items that would be clearly

[1] In contrast to latent-growth models, the mean structure that is expected to reflect the individuals' trajectories (i.e., individual differences) is not part of these models. The fixed-links modeling approach differs from item-response models on the item-position effect, which focus on the dependence of item difficulty parameters on the item position (Debeer & Janssen, 2013; Lozano & Revuelta, 2020), while the dependence of item discrimination parameters on item position has rarely been investigated (but see Nagy, Nagengast, Becker, Rose, & Frey, 2018).

dissociable from the item-position effect across the first 18 items.

More specifically, in both conditions, we assumed a latent variable representing reasoning ability. In the continuous rule condition, an item-position effect from the first to the 24th item was expected. Therefore, a bifactor model with two latent variables should describe the data well. In the discontinuous rule condition, the sudden change of rules should disrupt the item-position effect after the 18th item, requiring a *third* latent variable to appropriately describe the data. In other words, we expected a latent variable representing reasoning ability, a latent variable representing the item-position effect from the first to the 18th item and a third latent variable reflecting a new item-position effect across the last six items in the discontinuous rule condition.

2. Method

2.1. Participants

The experiment was conducted online using the QuestBack survey tool (EFS Survey, 2019). Undergraduates in psychology received course credit for their participation and all other participants could take part in a raffle for ten vouchers worth 20 CHF. The link to the study was accessed 1889 times, but 1264 individuals abandoned the experiment after accessing the welcome page. An automatic randomizer was programmed to randomly assign participants to the continuous rule (n = 216) or to the discontinuous rule condition (n = 213) after reaching the 18th item. Data of 429 individuals, who completed all items, were included in further analyses. Their mean age was 23.9 years (SD = 6.6). One person chose not to declare gender, while 149 reported to be male and 279 to be female. Fifty-nine participants reported a Bachelor's degree or higher as their highest level of education, 320 a university entrance qualification, and 48 participants had neither. All participants gave written informed consent prior to their participation by ticking a box on the welcome page of the survey. The study protocol was approved by the ethics committee of the Faculty of Human Sciences of the University of Bern (No. 2021–02-00003).

2.2. Test of figural analogies

Two Tests of Figural Analogies (TFA) were created using the IMak package in R developed by Blum and Holling (2018). Previous studies have demonstrated that figural analogy items generated with the IMak package can have satisfactory reliability and convergent validity with other analogy tests (Blum & Holling, 2018) although, these indicators cannot be transferred to newly created test items.

In this study, IMak items can follow one to three rules and are composed of a top and a bottom area (see Fig. 1). The top area consists of a 2 × 2 matrix with three figures and a question mark as the lower right entry. The bottom area of each item contains eight figures and two verbal responses ("No response is correct" and "I don't know") as response alternatives.

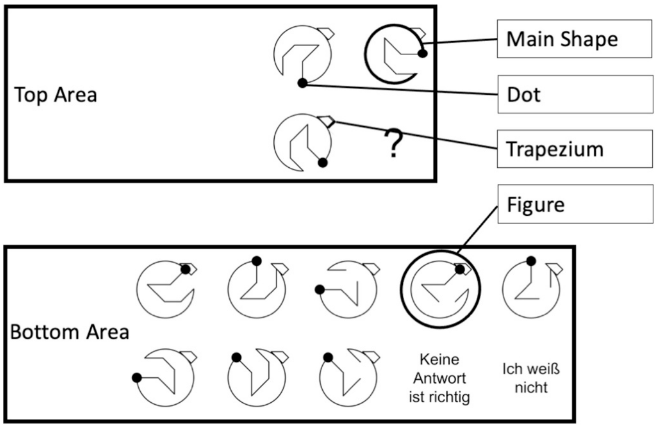


Fig. 1. Example item from the Test of Figural Analogies (TFA).
Note. Item created with the IMak package in R. Top left item in the bottom area is correct. "Keine Antwort ist richtig" means "No response is correct"; "Ich weiss nicht" means "I don't know".

To solve an item correctly a rule in the figures of the top area must be identified and applied to select the correct figure from the bottom area to substitute the question mark in the top area. The rule can either be deduced using the information provided by the two figures in the first row or column.

Each figure was composed of the elements "main shape", "dot" and "trapezium". The IMak package provides three main rules for item creation. The figure itself can be mirrored (Rule 1), straight lines of the main shape can be removed (Rule 2) or single elements can be moved (Rule 3). Within Rule 3, it is possible to move the trapezium or the opening of the figure clockwise and counterclockwise. It is also possible to choose the degree of movement, for example a 45° movement. The dot can be moved along the edges of the opening. Again, direction of movement and the number of edges the dot passes can be defined.

Instructions were translated and adapted from the supplementary material provided in Blum and Holling (2018). Participants first received a general instruction about the elements in an item (similar to Fig. 1), and a generic instruction on how to solve an item. The generic instruction emphasized, that the relation between the two figures in the first row (or the first column) should be identified and transferred to the second row (or column) to infer what the missing figure (represented by the question mark) should look like. The task was then to choose the corresponding figure from the bottom area. In this generic instruction no reference was made to any specific rule. The generic instruction also stated that, if participants could not find the corresponding solution in the figures of the bottom area, they should select the option "No response is correct" and if the item was too difficult for them to solve, to select the option "I don't know". This was followed by the same three practice items in both conditions. The three practice items only included Rule 3, moving the elements. For the practice items, the generic instruction was augmented with specific information about Rule 3 (e.g., direction of movement, what element was moved). Rule 3 was first applied to the main shape, then the trapezium, followed by the dot. Practice items had to be solved correctly to proceed with the test and were not included in any analyses reported below.

All items of the TFA version presented in the continuous rule condition were created using only Rule 3, where main shape, dot, or trapezium were moved (see Supplementary Material for example items). For the first six items, participants had to apply the rule only for one of the elements to solve it correctly. In the following nine items, two elements moved concurrently. The movements of the two elements were independent from each other. For example, the trapezium would move 45° clockwise while the main shape moved 90° counterclockwise. Finally, all three elements (i.e., main shape, dot, and trapezium) moved simultaneously in the last nine items. Again, the degree and direction of the movement was different for each element. This ensured that the rules applied did not accidentally cancel each other out.

In the TFA version used for the discontinuous rule condition, the first 18 items were identical to the TFA version in the continuous rule condition. Thereafter, six items were created with the remaining two rules. For each item, the figure itself had to be mirrored (Rule 1) and the correct line of the main shape had to be removed (Rule 2) to identify the correct answer (see Supplementary Material for example items - contact corresponding author for full TFA item sets). No time limit was set in either condition. Participants only received feedback for practice items.

2.3. Procedure

Participants were informed prior to their participation that mobile devices would not work. If a participant accessed the study with a mobile device, the study was automatically terminated after the welcome page. This ensured a reasonable screen size and similar handling for giving responses across all participants. After the welcome page with information about the study and the request to confirm their informed consent participants responded to several demographic questions (participants indicating a diagnosed learning disability were excluded from the sample), followed by the TFA. The TFA version was randomly assigned upon completion of the 18th item. Thereafter, participants completed further tasks irrelevant for the current study.[2] On the final page all participants were given contact information for potential questions and were informed about the purpose of the study.

2.4. *Statistical analysis*

Analyses were run with R software using the *lavaan* (Rosseel, 2012) and *psych* (Revelle, 2011) packages. To compare participants' test scores between conditions, two sample Welch tests were calculated. To test the main hypotheses different one- and bifactor models were fit to the data.

All CFAs were based on probability-based covariance matrices[3] as suggested by Schweizer (2013) for binary data and robust maximum likelihood estimation. To account for the difference between binary data distribution and normal distribution assumed for the latent constructs, the factor loadings were weighted by the standard deviation of the respective item (Schweizer, 2013).

For the investigation of the item-position effect, the continuous rule condition served as the control or reference condition as we expected the item-position effect to develop across all 24 items. Therefore, we first analyzed the continuous rule condition data for it's fit regarding three different models (see Table 1). The one-factor model (Model 1) was compared with two bifactor models (illustrated in Fig. 2, Panel A) to examine whether the consideration of an item-position effect as a second latent variable improved data description.

The factor loadings in the bifactor models of the latent variable reflecting the item-position effect were either fixed to linearly (Model 2) or quadratically increase (Model 3) from the first to the last item. These two different courses are most often investigated (e.g., Schweizer & Troche, 2018; Schweizer, Troche, & Rammsayer, 2011). The correlation between the latent variables representing reasoning ability and the item-position effect was set to zero. All variances of latent variables with fixed factor loadings were estimated freely. Since the size of these variances (and their standard errors) depends on the height of the chosen factor loadings, the variances of the final model were scaled according to the eigenvalue scaling method (Schweizer & Troche, 2018). Comparing the scaled variances provides information about the relative strength of the latent variables within the model.

Table 1
Overview of calculated models and details regarding the specifications.

Model	Structure	Latent variable(s)	Fixation of factor loadings
Model 1	One-factor model	Reasoning ability	1
Model 2	Bifactor model	Reasoning ability Item-position effect (linear course)	1 $\frac{i}{k}$
Model 3	Bifactor model	Reasoning ability Item-position effect (quadratic course)	1 $\frac{i^2}{k^2}$

Note. All factor loadings were additionally weighted by SD_i as link function. Letter *i* refers to the respective items position, *k* refers to the total number of items (i.e., 24), SD to the standard deviation.

The same three models (Table 1) were fit to the 24 TFA items in the discontinuous rule condition. A bifactor model was expected to result in a good fit for the continuous rule condition but an inadequate fit for the discontinuous rule condition due to the experimental manipulation of rules.

Therefore, in a final step, additional bifactor models with *three* latent variables (see Fig. 2, Panel B) were fit to the data of each condition. In these models, the latent variable representing the item-position effect was modeled with increasing factor loadings from the first to the 18th item. For the last six items (19 to 24) a third latent variable was defined. The factor loadings continued the course set for the latent variable representing the item-position effect. Again, the correlations between the latent variables were set to zero. Further models where, for example, factor loadings started anew for the third latent variable were also calculated to examine alternative explanations (see Supplementary Analyses).

All models were evaluated regarding their fit indices as recommended by DiStefano (2016). For the Root Mean Squared Error of Approximation (RMSEA), values below 0.08 indicate an acceptable and below 0.06 a good fit; Standardized Root Mean Square Residual (SRMR) indicates an acceptable fit below 0.10 and a good fit below 0.08; Comparative Fit Index (CFI) and Tucker Lewis index (TLI) are evaluated as acceptable with values above 0.90 and as good with values above 0.95. Model comparisons were based on comparisons of CFI, the Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Lower values for AIC and BIC indicate better fit and higher values of at least 0.010 in the CFI indicate better fit, respectively (Chen, 2007).

Additionally, all latent variables in the model should have a positive and statistically significant variance. If a variance parameter is not statistically significant, there is no added value by that latent variable, and there is no reason to keep it in a model. The variance is tested for significance one-tailed, since negative values are theoretically impossible for variance parameters and indicate grave misspecification of a model.

For the final models, the omega coefficients of the latent variables were computed according to Bollen (1980). Data and code for the analysis can be found in the Supplementary Materials.

3. Results

Since the TFA was completed online and no supervision as in a laboratory setting was possible, two quality checks were put in place. We excluded all participants that took <6 min to complete the TFA to ensure that only data of participants was included who completed the test with enough diligence. From the total of 429 participants who completed all TFA items, 16 participants were excluded. We also excluded another 10 participants since they took longer than an hour to complete all items.

Overall, 403 individuals passed the quality checks retaining a balanced allocation between the continuous rule (n = 203) and discontinuous rule (n = 200) condition. In the continuous rule condition, the mean age of participants was 25.9 years (SD = 6.6), and 68 reported their sex as male and 135 as female. Twenty-eight participants reported a Bachelor's degree or higher as their highest level of education, 151 had a university entrance qualification, and 24 participants had neither. In the discontinuous rule condition, the mean age of participants was 25.1 years (SD = 5.6), and 63 reported their sex as male, 136 as female, and one person chose not to declare. Twenty-nine participants reported a Bachelor's degree or higher as their highest level of education, 156 had a university entrance qualification, and 15 participants had neither.

Descriptive statistics for test scores in the two TFA versions for the continuous and the discontinuous rule condition are presented in Table 2. Mean accuracy and standard deviations for the 24 items of each TFA version are presented in Fig. 3. Test scores of the first 18 items were very similar across conditions, $t(400.96) = 0.881$, $p = .379$, $d = 0.088$.

[2] Solving-strategy questionnaire (8 items), Framed-Line Test (20 Items), Figural Matrices (30 Items), Solving-strategy questionnaire (8 items), 5 items regarding the online experience while completing the study, short written debriefing and contact information of responsible investigator.

[3] Models based on tetrachoric correlations, with weighted least squares estimator yield similar results and are reported in the Supplementary Materials (Table E).

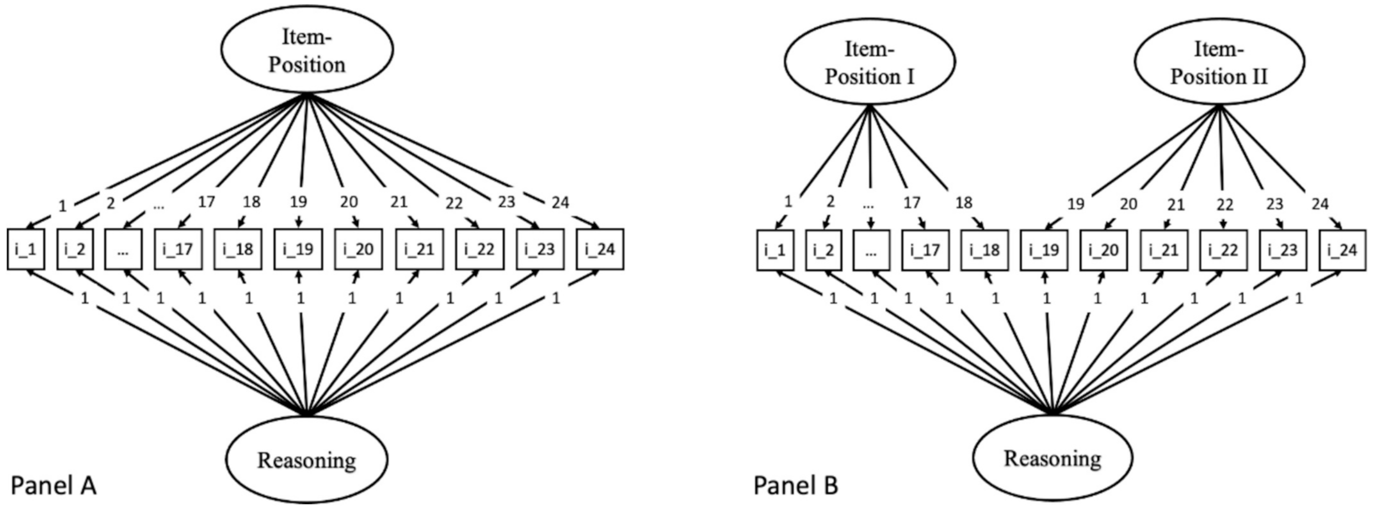


Fig. 2. Bifactor models with two or three latent variables.
Note. Fixations of the factor loadings of the item-position effect are either directly divided by k (total number of items, 24) for a linear increase or first squared and then divided by k^2 (total number of items squared, 24^2) for a quadratic increase. Panel A shows a bifactor model with *two* latent variables. Panel B a bifactor model with *three* latent variables. For both models covariances and means of latent variables were set to zero.

Table 2
Descriptive test statistics for the figural analogies test (TFA).

Condition	Items	Mean	SD	Min	Max	Skewness	Kurtosis	Cronbach's α
Continuous rule ($n = 203$)	1–18	13.46	3.58	3	18	−0.81	0.04	0.81
	1–24	17.43	5.05	3	24	−0.75	−0.29	0.86
Discontinuous rule ($n = 200$)	1–18	13.15	3.56	2	18	−0.90	0.42	0.80
	1–24	15.81	4.73	2	24	−0.51	−0.08	0.83

Note. For each condition the first 18 items (1–18) and the full set of 24 items (1–24) with their respective descriptive statistics are presented.

This is also visible in the mean accuracies displayed in Fig. 3. When all 24 items were compared, the performance in the continuous rule condition was significantly better than in the discontinuous rule condition, t (400.02) $= 3.342$, $p < .001$, $d = 0.333$. As visible in Fig. 3, the introduction of a new rule in the discontinuous rule condition increased item-difficulty. This could explain the difference in performance between conditions. The comparison of all 24 items is informative but needs cautious interpretation since it is based on different items. Participants in both conditions indicated similar diligence for completing the TFA, t (380.32) $= 0.631$, $p = .529$, $d = 0.063$. Participants strongly agreed when asked whether they took their time to inspect each item to think of the correct solution, continuous rule condition = 7.47 (SD = 1.99), discontinuous rule condition = 7.58 (SD = 1.55) on a 9-point rating scale. In both conditions, the TFA showed good internal consistency in terms of Cronbach's alpha (see Table 2). This was true regardless of whether the first 18 items were examined, or all 24 items.

In the first step of investigating the item-position effect, the focus was on the continuous rule condition, where the movement rule was used across all 24 items. As can be taken from the upper part of Table 3, the one-factor model (Model 1) led to an adequate data description according to CFI, TLI, and SRMR (see Table 3). However, when a latent variable reflecting the item-position effect was added (Model 2 & 3), model fit improved. Specifically, both bifactor models clearly had smaller AIC and BIC values, higher CFI and TLI values and smaller RMSEA and SRMR values compared to Model 1. The CFI difference exceeded the threshold of 0.010, which is considered to reflect a substantial difference (Chen, 2007). In both bifactor models, the variance parameters of the latent variables yielded statistical significance indicating that there was a substantial portion of variance being explained by both latent variables. The overall fit of Model 2 and Model 3 was very similar. Nevertheless, Model 2 described the data somewhat better according to AIC and BIC. Regarding CFI, TLI, RMSEA, and SRMR the differences were rather small but always preferential for Model 2.

Summarized, a linear item-position effect across the 24 items in the continuous rule condition could be identified since Model 2 provided the best data description with a latent variable reflecting reasoning ability and a second latent variable describing a linearly increasing item-position effect. The scaled variance parameters were $\varphi = 0.699$ ($z = 5.960$, $p < .001$) for latent reasoning ability and $\varphi = 0.342$ ($z = 4.542$, $p < .001$) for the item-position effect indicating that about a third of the latent variance could be attributed to the item-position effect. Omega coefficients were $\Omega = 0.83$ and $\Omega = 0.64$, respectively.

In the next step, we investigated the same three models in the discontinuous rule condition (see Table 3). The overall pattern of results was somewhat similar to the continuous rule condition: The bifactor models (Models 2 & 3) described the data better than the one-factor model (Model 1) according to all fit indices except for SRMR. The variance parameters of the second latent variable reflecting the item-position effect in the bifactor models were statistically significant. However, while Model 2 and 3 provided a better data description than Model 1, the fit indices (and primarily CFI and TLI) indicated that the description was inadequate.

It is plausible to assume that the experimental manipulation of the rules caused the different results between conditions. Upon the 18th item, scores and completion time were almost identical. With the experimental manipulation, the response behavior for the last six items in the discontinuous rule condition was expected to differ hence a third latent variable might be required in the discontinuous rule condition.

Therefore, two additional bifactor models with *three* latent variables were calculated for the discontinuous rule condition (see Panel B of Fig. 2). Proceeding from the results in the continuous rule condition, the factor loadings of the latent variable reflecting the item-position effects were modeled to increase linearly (as in Model 2). In the first model (Model 2a), the correlations between all latent variables were set to zero.

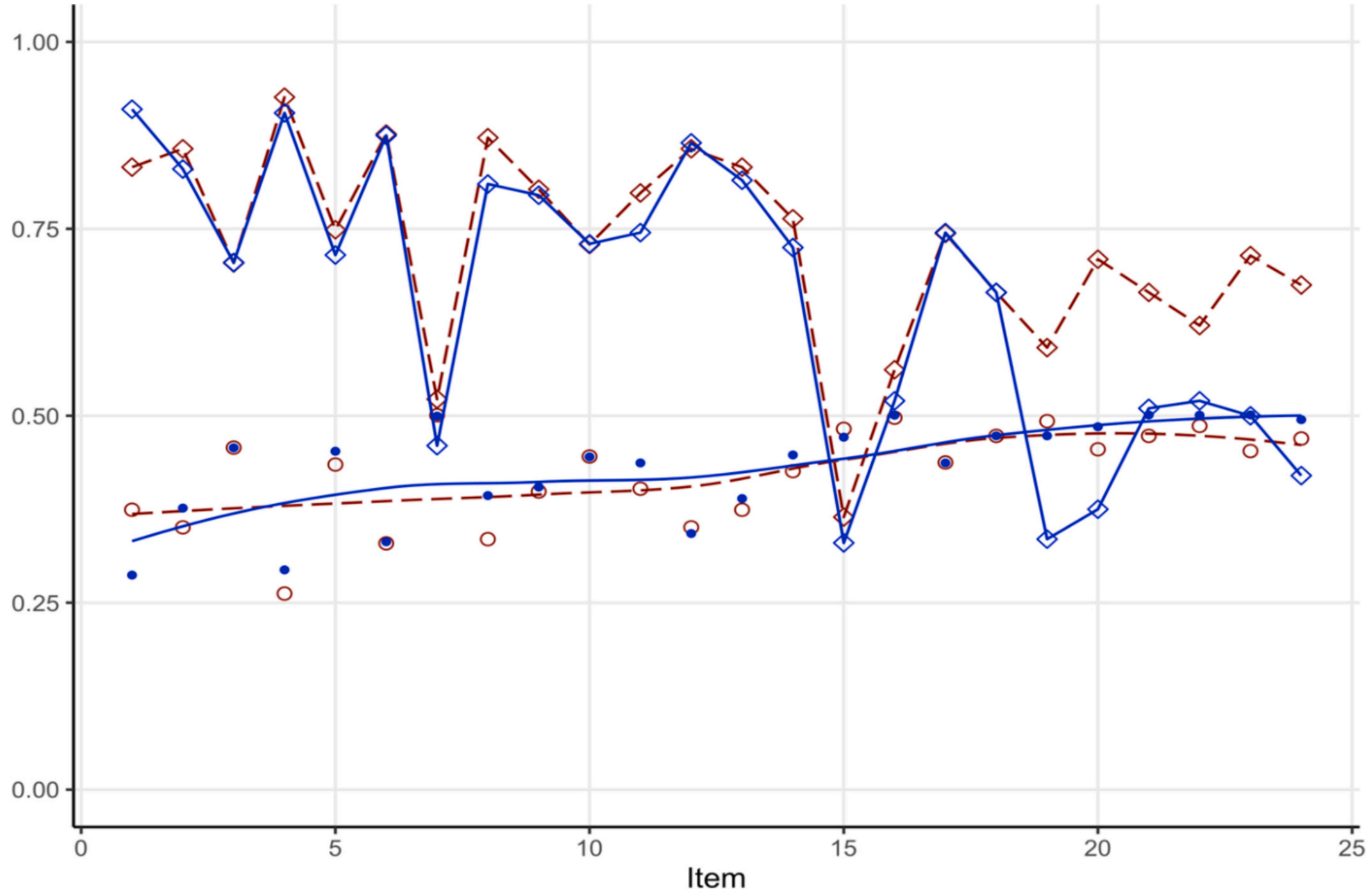


Fig. 3. Accuracy and standard deviation for each item in both conditions.
Note. Dashed lines depict values in the continuous rule condition, solid lines in the discontinuous rule condition. Squares refer to item accuracy, empty and filled circles to the standard deviation in the continuous and in the discontinuous rule condition, respectively. Smooth lines depict the trend lines of the standard deviations in the two conditions.

Table 3
Measurement models with one or two latent variables calculated for both TFA versions (continuous rule and the discontinuous rule condition).

Condition	χ^2 *(df)*	*p*	AIC	BIC	CFI	TLI	RMSEA	SRMR	*p* ($\varphi_{\text{reasoning}}$)	*p* ($\varphi_{\text{IPE 1}}$)
Continuous rule										
Model 1: one-factor model	314.81 *(275)*	0.049	4636	4718	0.940	0.940	0.029	0.089	<0.001	
Model 2: bifactor model: IPE fixed to linearly increase	286.79 *(274)*	0.286	4605	4691	0.981	0.981	0.017	0.076	<0.001	<0.001
Model 3: bifactor model: IPE fixed to quadratically increase	289.65 *(274)*	0.247	4608	4694	0.977	0.976	0.018	0.078	<0.001	<0.001
Discontinuous rule										
Model 1: one-factor model	458.61 *(275)*	<0.001	4981	5063	0.731	0.730	0.060	0.096	<0.001	
Model 2: bifactor model: IPE fixed to linearly increase	412.09 *(274)*	<0.001	4937	5023	0.796	0.794	0.052	0.094	<0.001	<0.001
Model 3: bifactor model: IPE fixed to quadratically increase	382.26 *(274)*	<0.001	4907	4993	0.839	0.838	0.046	0.096	<0.001	<0.001

In the second model (Model 2b) the correlation between the two variables reflecting the item-position effects across the first eighteen and the last six items was estimated (see Table 4). For the discontinuous rule condition, both models (Model 2a and 2b) led to an adequate data description according to CFI and TLI as well as a good data description according to RMSEA and SRMR. Additionally, variance parameters of all three latent variables were statistically significant. Therefore, in contrast to the continuous rule condition, three instead of two latent variables were necessary to explain individual differences of response behavior in the discontinuous rule condition. This becomes most evident when comparing the model fit for the discontinuous rule condition in Table 3 and Table 4.

Model 2a and 2b described the data similarly well according to CFI, TLI, RMSEA, and SRMR. This was due to the correlation between the two item-position effects not being statistically significant when estimated freely, $r = 0.173, p = .378$. This supported the idea that the item-position effect extracted from the last six items with new rules was independent from the item-position effect across the first 18 items. Consequently, the

Table 4
Bifactor models with three latent variables where the correlation between the two item-position effect latent variables was set to zero or estimated freely.

Condition	χ^2 *(df)*	*p*	AIC	BIC	CFI	TLI	RMSEA	SRMR	*p* (φ $_{\text{reasoning}}$)	*p* (φ $_{\text{IPE 1}}$)	*p* (φ $_{\text{IPE 2}}$)
Continuous rule											
Model 2a / three latent variables, all correlations set to zero	293.56 *(273)*	0.188	4613	4703	0.969	0.969	0.021	0.083	<0.001	0.072	<0.001
Model 2b / three latent variables, freely estimated correlation between the two IPEs	280.23 *(272)*	0.353	4600	4693	0.988	0.987	0.013	0.076	<0.001	0.001	<0.001
Discontinuous rule											
Model 2a / three latent variables, all correlations set to zero	321.85 *(273)*	0.022	4845	4934	0.927	0.927	0.031	0.078	<0.001	0.001	<0.001
Model 2b / three latent variables, freely estimated correlation between the two IPEs	320.80 *(272)*	0.022	4846	4938	0.927	0.926	0.031	0.078	<0.001	0.002	<0.001

Note. Given are chi-square (χ^2) values with degrees of freedom (df), the respective *p*-value and the fit indices, RMSEA, SRMR, CFI, and AIC. IPE = item-position effect. Last three columns show the *p*-value for the respective latent variable, indicating whether it described a significant portion of variance.

fit of Model 2a and 2b was similar but Model 2a provided a slightly more parsimonious data description and therefore should be preferred. In Model 2a, the scaled variance parameters of the latent variables were φ = 0.633 (z = 5.561, p < .001) for reasoning ability, φ = 0.144 (z = 3.026, p = .002) for the item-position effect across the first 18 items, and φ = 0.370 (z = 7.153, p < .001) for the item-position effect across the last six items. Omega coefficient of the reasoning latent variable was Ω = 0.81, while the omega coefficients for the first and the second item-position effect latent variable were Ω = 0.43 and Ω = 0.70, respectively.

We also tested whether these bifactor models with *three* latent variables would better fit the data in the continuous rule condition (see Table 4). With the first latent variable depicting the item-position effect in Model 2a not explaining a significant amount of variance, the model must be rejected despite acceptable fit. In Model 2b, where the correlation between the two item-position effects was freely estimated, all three latent variables explained a significant amount of variance (reasoning ability: φ = 0.680, z = 5.619, p < .001; first item-position effect: within the first 18 items: φ = 0.185 z = 3.137, p = .002; second item-position effect: φ = 0.223, z = 4.624, p < .001). Interestingly, the model fit was similar to the bifactor model with only two factors (Model 2). This can be explained by the large correlation between the two latent variables representing item-position effects, r = 0.720, p = .001. Such a high correlational relationship points to a substantial overlap between the first and the second item-position effect and casts doubt on the need for an additional item-position effect for the last six items.

Summarized, the bifactor model with two latent variables (Model 2) provided a good data description in the continuous, but not in the discontinuous rule condition. In the discontinuous rule condition, the bifactor model with *three* latent variables (Model 2a) provided a substantially better data description. Hence, due to the experimental manipulation no configural invariance could be obtained between the two TFA versions of the continuous and discontinuous rule condition. When the factorial structure of a test does not show configural invariance between groups, stricter comparisons (e.g., for metric or scalar invariance) are not appropriate.

3.1. Supplementary analyses

For a comprehensive analysis and a closer look at possible alternative explanations, additional models were calculated. Here, we give a brief overview; detailed information on the models can be found in the Supplementary Material.

A frequent assumption is that the item-position effect reflects increasing item difficulty of items in reasoning tests, although several studies ruled this out (e.g., Schweizer & Troche, 2018; Zeller, Reiss, & Schweizer, 2017). Nonetheless, additional models where item difficulty was used to fix the factor loadings of the latent variable depicting the item-position effect were calculated. Results are presented in the Supplementary Material (Table A). Data description was worse for the continuous rule condition. For the discontinuous rule condition, the bifactor model with three latent variables in Table 4 outperformed the bifactor model considering item difficulty. Therefore, we conclude that the item-position effect is not a reflection of item difficulty (for similar results, see Schweizer & Troche, 2018; Zeller, Reiss, & Schweizer, 2017).

An alternative idea is that the number of rules underlie the item-position effect rather than item position. This can be addressed with the current experimental design since the number of rules differs between conditions for the last six items. The bifactor models with two latent variables were calculated again but the factor loadings of the latent variable representing the item-position effect were fixed according to the number of rules underlying the respective item. Results are reported in the Supplementary Material (Table B) and indicated no noteworthy change in model fit for the continuous rule condition. Given the linear increase of rules in this condition, this result is not surprising. In the discontinuous rule condition, this approach described the data worse than the bifactor model with a linear or quadratic increase of factor loadings for the item-position effect. Although our study was not designed to specifically test this assumption, our results suggest that it is unlikely that the item-position effect reflects the increasing number of rules used in the items.

An anonymous reviewer raised the concern, that some sort of speed component could underly the item-position effect. To test this hypothesis, we ran additional models with median response latencies as constraints. From the results (Table C in Supplementary Material), we conclude that the consideration of response latencies leads to similar (but somewhat worse) model descriptions in both conditions. Importantly, the qualitative differences between the first 18 and the last six items in the discontinuous rule condition regarding the systematic change of rules are nonetheless visible. If a change in processing speed (and individual differences therein) accounted for the break in the item-position effect, the consideration of this change should lead to a common factor (or at least to correlated factors). This was not the case.

For the bifactor models with *three* latent variables, the pattern of results did not depend on a specific way of fixing the factor loadings of the third latent variable. Here, we used the course of 19 to 24 as constraints of factor loadings for the last six items, to keep the models as comparable as possible. Model fit and correlations somewhat changed when these constraints were set to increase from 1to 6 in order to depict a new start of the additional item-position effect (for details see Table D in the Supplementary Materials). However, the overall pattern of results remained the same.

Additionally, an anonymous reviewer raised the question what the model fit would be, if the factor loadings for the additional third latent variable were estimated freely. This led to a slightly better model fit, $\chi^2(272) = 317.824$, AIC = 4842.829, BIC = 4935.182, CFI = 0.931, TLI = 0.930, RMSEA = 0.029, SRMR = 0.078, compared to the model fit reported in Table 4. Interestingly, the factor loadings of items 19 to 24 were very similar, varying between 0.472 and 0.500. The correlation of these factor loadings with the factor loadings depicting a linear increase is very high ($r = 0.74$), albeit questionable due to the small number of data points correlated.

4. Discussion

Although reasoning ability tests use very homogenous item material, the demonstration of unidimensionality has been frequently challenged. Bifactor models, considering a second latent variable in addition to reasoning ability to capture increasing (true) variance from the first to the last item often led to a substantially better data description. Since the factor loadings of this second latent variable reflect the position of the respective item in a test, this latent variable is a representation of the item-position effect. The origin of the increase in variance is an ongoing debate (e.g., Birney et al., 2017; Embretson, 1995; Lozano & Revuelta, 2020; Ren et al., 2014). The aim of this experiment was to test the learning hypothesis (Ren et al., 2014) stating that the item-position effect can be explained by the learning of rules during test taking.

Consistent with our expectations and as reported in several other studies (e.g., Ren, Gong, et al., 2017), a typical item-position effect could be identified in the continuous rule condition, since a bifactor model with two latent variables yielded the best model/data fit. For the discontinuous rule condition, the analog bifactor model did not lead to an acceptable data description. Fitting a quadratic instead of a linear course of factor loadings for the latent item-position variable, provided only a marginally better model fit. Thus, an acceleration of the item-position effect in the discontinuous compared to the continuous rule condition can be ruled out. A model with three latent variables (Model 2a/b) resulted in an overall better data description in the discontinuous rule condition.

In the bifactor models with three latent variables, the two latent variables depicting item-position effects were independent of each other in the discontinuous rule condition but closely related in the continuous rule condition. Thus, when the rules underlying the test items were the same (as in the continuous rule condition), the two latent variables correlated highly, and a single latent variable for the item-position effect described the data just as adequately as two highly correlated latent variables. However, the correlation between the two item-position effects fell to a non-significant level when the rules were changed in the discontinuous rule condition. Hence only in the continuous rule condition, the item-position effect continued across the last six items where the rules remained unchanged, contrary to the discontinuous condition where new rules were used. This indicates that the item-position effect depended on repeating the same rule rather than simple (increasing) familiarity with the stimulus material (which was the same across all 24 items in both conditions) or the length of the test and possible fatigue effects.

Summarized, results suggest that repeatedly being confronted with the same rule led to individual differences in the response behavior as represented by the latent item-position variable. This means that some individuals gained a larger advantage from rule repetition than others. In combination with previous results on the functional correlational relationship between the item-position effect and rule learning in an external task (Ren et al., 2014; Schweizer et al., 2021), the present findings provide further evidence for the learning hypothesis (Ren et al., 2014).

Certainly, there are other explanations for the present findings. We tried to rule out some in the analyses reported in the supplementary material. It is common practice to arrange items of a reasoning test according to their difficulty to avoid discouraging participants at the beginning. We chose the same procedure, and later items were more difficult than earlier items (see Fig. 2). Alas, item position cannot be clearly dissociated from item difficulty in either condition of the current experiment. Yet, if the item-position effect was due to item difficulty, the data description should be best when the factor loadings were fixed according to item difficulty. This was not the case for the continuous rule condition (see Table A in supplementary materials). For the discontinuous rule condition, there was a slight improvement in fit, but the bifactor model with three independent latent variables (in Table 4) still outperformed the bifactor model considering item difficulty. Therefore, we conclude that there is a qualitative difference between the first eighteen and the last six items in the discontinuous rule condition and not a quantitative graduation (as one would expect for item difficulty). Based on these results, it is unlikely that item difficulty is the source of the item-position effect, which is consistent with previous studies (Birney et al., 2017; Schweizer & Troche, 2018; Zeller, Reiss, & Schweizer, 2017).

It is important to distinguish between item difficulty and item complexity. For example, Spilsbury, Stankov, and Roberts (1990) showed that the number of rules refer to item complexity but not necessarily to item difficulty. Carpenter et al. (1990) argued that items with more rules are more complex and that an increase in complexity leads to an increase in working memory load. This phenomenon should lead to an increase in variance throughout the test since individual differences in working memory capacity would have a greater impact on the later items with more rules than on the earlier items with fewer rules (see also Embretson, 1995).

Assuming that item complexity is reflected in the number of rules underlying an item, the design of the current experiment allows for further exploration of this idea. While in the continuous rule condition the number of rules and item position increase (more or less) simultaneously, this is not the case in the discontinuous rule condition. Here only two rules had to be applied in the last six items but three rules in the six items before, allowing for a separation of item position and item complexity. Constraining the factor loadings of the second latent variable according to the number of rules did not lead to a better data description than the bifactor models with three latent variables in the discontinuous rule condition (see Table B in supplement). Therefore, it seems unlikely that item complexity (i.e., number of rules) accounts for the item-position effect.

These conclusions were further corroborated by the additional analyses using median response latencies to fix the factor loadings of the latent variable reflecting the item-position effect. Response latencies in reasoning tests often increase with increasing item difficulty and/or complexity (e.g., Neubauer, 1990; Vigneau, Caissie, & Bors, 2006). Hence, if the item-position effect reflects task difficulty or complexity, the fixation of factor loadings according to median response latencies should have provided a good model fit for both conditions regardless of the experimental manipulation. This could not be confirmed by our analyses making item difficulty and complexity unlikely candidates for the underlying source of the item-position effect.

Nagy, Ulitzsch, and Lindner (2023) emphasized the importance of rapid guessing or disengagement for understanding changes in response behavior during test taking. We asked participants about their test-taking experience since the experiment was conducted online. When asked if they took their time to inspect the item material and to think about the correct solution, 75% of the participants' self-reports indicated a value of 7 or higher (on 9-point rating scale). These self-reports did not differ between condition. Therefore, it seems unlikely that the different factor structure of the TFA between the two conditions (configural variance) would be due to careless test-taking behavior. Nevertheless, self-reports are far from objective. Individual response latencies (and possible interactions with self-reports; see Nagy et al., 2023) might provide more objective information about rapid guessing. With the current data we cannot confidently rule out that the sudden change of rules led to individuals guessing the answers or disengaging. Such behavior could indeed explain the drop in accuracy in the discontinuous rule condition and hence present an alternative explanation for the current results.

Future research should implement useful controls regarding disengagement and rapid guessing. For example, tracking participants' eye-movements could provide more detailed and objective information about response behavior and possibly reveal participants disengaging or guessing. Furthermore, previous investigations of eye-movements during test taking uncovered that participants apply different strategies to solve items (e.g., Laurence, Mecca, Serpa, Martin, & Macedo, 2018). A possible relationship between individual differences, the applied strategiy and the item-position effect has not yet been investigated. A sudden change in rules might elicit a change in strategy. This could serve as an alternative explanation for the present results and should be addressed in future research.

Nonetheless, if we cautiously maintain that current results support the learning hypothesis on the item-position effect, follow-up questions arise. For example, it would be important to learn more about the type of learning reflected in the item-position effect. In experimental tasks of two previous studies, participants were explicitly instructed to learn given rules (Ren et al., 2014; Schweizer et al., 2021). Individuals who learned these rules more successfully also had a more pronounced item-position effect. These findings suggest that the item-position effect most likely reflects some sort of explicit learning. The influence of the explicit learning of rules prior to the test (e.g., Loesche, Wiley, & Hasselhorn, 2015) or feedback about the relevant rules after each item (e.g., Guthke & Stein, 1996) on the item-position effect might provide further information about its characteristics and nature. It would also be interesting to examine whether the disruption of the item-position effect due to the sudden rule change would still emerge if the new rules were explained at the beginning and were part of the practice items.

To date, no study has addressed the relation of the item-position effect to other types of learning such as implicit learning. The item-position effect is usually unrelated to reasoning ability even when this correlation is freely estimated (Schweizer et al., 2021). This was also true in the present study, where the correlation between reasoning ability and the item-position effect in the continuous rule condition was $r = 0.055$, $p = .806$, when freely estimated. Such an independence from reasoning ability was also reported for measures of implicit learning (e.g., Danner, Hagemann, & Funke, 2017; Kalra, Gabrieli, & Finn, 2019). Therefore, it might be an interesting avenue for future studies to better understand whether the item-position effect relates to implicit learning. Such investigations would also help improve the largely unexplored understanding of *where* in structural models of intelligence the ability underlying the item-position effect could be located.

The counterintuitive finding that reasoning ability and learning (if it is the construct underlying the item-position effect) are unrelated points to a complex interaction. From a conceptual perspective, it seems plausible that individuals with higher reasoning ability gain more from (successfully) solving earlier items to solve later items compared to individuals with lower reasoning abilities. This would imply a positive and linear relation between reasoning and the item-position effect. Yet, for individuals with very high reasoning ability, the learning of rules might become redundant after the first few items and hence would result in a very faint item-position effect. Similarly, individuals with very low reasoning abilities will probably have difficulties understanding the rules and will not be able learn them by (unsuccessfully) solving items. Hence, a faint item-position effect can be associated with very high or very low reasoning abilities. Thus, the relationship between reasoning ability and the item-position effect might not be linear. These assumptions are highly speculative in nature and definitely require further investigation with large subsamples of individuals with varying reasoning abilities.

Proactive interference might also be at play when analyzing individual differences in the item-position effect and rule learning behavior. Proactive interference describes the phenomenon that information that was previously encoded competes with current information while it is no longer relevant (Hamilton, Ross, Blaser, & Kaldy, 2022). The ability to disengage from no longer relevant information has been identified as a major limitation on working memory capacity (Oberauer, Farrell, Jarrold, & Lewandowsky, 2016). This kind of disengagement has been proposed to be especially beneficial in reasoning ability tests to block the retrieval of faulty hypotheses about the underlying rules of an item (Burgoyne & Engle, 2020; Shipstead, Harrison, & Engle, 2016). That such mechanisms of attention control can add to the understanding of the item-position effect has been demonstrated by von Gugelberg et al. (2021). Their results indicate that individuals who predominately engaged in proactive cognitive control showed a stronger item-position effect than individuals using reactive cognitive control.

A major limitation of the present study is the lack of external variables such as indicators of working memory capacity or proactive interference. They could have provided valuable information about the nature of the item-position effect and shed some light on the differences between the first and second item-position effect in the discontinuous rule condition (Model 2b).

Moreover, for the second item-position effect in the discontinuous rule condition, no clear increase of the factor loadings could be detected when factor loadings were estimated freely casting doubt on the assumption of a second item-position effect. It is possible that there were simply too few items for a new learning effect to develop and future research should address how many rule repetitions are necessary to elicit an item-position effect.

The goal of the present experiment was to better understand the item-position effect as it has been commonly investigated by means of fixed-links modeling. However, there are other methodological approaches to the item-position effect, namely item-response theoretical (Lozano & Revuelta, 2020) and multilevel-modeling approaches (Birney et al., 2017). Future research on the item-position effect could benefit from a better understanding of how these approaches differ and where they converge.

In sum, as recommended by Cronbach (1957), we introduced an experimental design (i.e., the comparison of the continuous and the discontinuous rule condition) allowing to test differences between correlation-based latent variable models in two experimental conditions. The current experiment delivered new tentative evidence for the learning hypothesis on the item-position effect. Due to the experimental manipulation of rules, configural invariance between the continuous rule and the discontinuous rule condition could not be obtained indicating that the repetition of rules seems to be a necessary precondition of the item-position effect. While the learning of rules seems a plausible explanation for the item-position effect, other explanations cannot be ruled out completely (e.g., rapid guessing, disengagement, item-solving strategy). Additional research is needed to provide more clarity on the subject and further determine how task properties (familiarity, rule novelty, etc.) influence the item-position effect.

CRediT authorship contribution statement

Helene M. von Gugelberg: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Karl Schweizer:** Writing – review & editing, Validation, Methodology. **Stefan J. Troche:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Conceptualization.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lindif.2024.102622.

References

Birney, D. P., Beckmann, J. F., Beckmann, N., & Double, K. S. (2017). Beyond the intellect: Complexity and learning trajectories in Raven's progressive matrices depend on self-regulatory processes and conative dispositions. *Intelligence, 61*, 63–77. https://doi.org/10.1016/j.intell.2017.01.005

Blum, D., & Holling, H. (2018). Automatic generation of figural analogies with the IMak package. *Frontiers in Psychology, 1286*. https://doi.org/10.3389/fpsyg.2018.01286

Bollen, K. A. (1980). Issues in the comparative measurement of political democracy. *American Sociological Review, 45*(3), 370–390. https://doi.org/10.2307/2095172

Burgoyne, A. P., & Engle, R. W. (2020). Attention control: A cornerstone of higher-order cognition. *Current Directions in Psychological Science, 29*(6), 624–630. https://doi.org/10.1177/096372142096937

Carlstedt, B., Gustafsson, J.-E., & Ullstadius, E. (2000). Item sequencing effects on the measurement of fluid intelligence. *Intelligence, 28*(2), 145–160. https://doi.org/10.1016/S0160-2896(00)00034-9

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven progressive matrices test. *Psychological Review, 97*(3), 404. https://doi.org/10.1037/0033-295X.97.3.404

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671.

Danner, D., Hagemann, D., & Funke, J. (2017). Measuring individual differences in implicit learning with artificial grammar learning tasks. *Zeitschrift für Psychologie, 225*(1), 5–19.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164–185. https://doi.org/10.1111/jedm.12009

DeShon, R. P., Chan, D., & Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's advanced progressive matrices: Evidence for multidimensional performance determinants. *Intelligence, 21*(2), 135–155. https://doi.org/10.1016/0160-2896(95)90023-3

Dillon, R. F., Pohlmann, J. T., & Lohman, D. F. (1981). A factor analysis of Raven's advanced progressive matrices freed of difficulty factors. *Educational and Psychological Measurement, 41*(4), 1295–1302. https://doi.org/10.1177/001316448104100438

DiStefano, C. (2016). Examining fit with structural equation models. In K. Schweizer, & C. DiStefano (Eds.), *Principles and Methods of Test Construction* (pp. 166–196). Hogrefe.

EFS Survey. (2019). Version EFS Winter 2018. Questback GmbH. Cologne: Questback GmbH. URL https://www.unipark.com/en/survey-software/ [accessed 2022-05-19].

Embretson, S. E. (1995). The role of working memory capacity and general control processes in intelligence. *Intelligence, 20*(2), 169–189. https://doi.org/10.1016/0160-2896(95)90031-4

Formann, A. K., Piswanger, K., & Waldherr, K. (2011). *Wiener Matrizen-Test 2: Ein Rasch-skalierter sprachfreier Kurztest zu Erfassung der Intelligenz*. Hogrefe.

Gijselaers, H. J., Meijs, C., Neroni, J., Kirschner, P. A., & de Groot, R. H. (2017). Updating and not shifting predicts learning performance in young and middle-aged adults. *Mind, Brain, and Education, 11*(4), 190–200. https://doi.org/10.1111/mbe.12147

Guthke, J., & Stein, H. (1996). Are learning tests the better version of intelligence tests? *European Journal of Psychological Assessment, 12*(1), 1–13. https://doi.org/10.1027/1015-5759.12.1.1

Hamilton, M., Ross, A., Blaser, E., & Kaldy, Z. (2022). Proactive interference and the development of working memory. *Wiley Interdisciplinary Reviews: Cognitive Science, 13*(3), Article e1593. https://doi.org/10.1002/wcs.1593

Harrison, T. L., Shipstead, Z., & Engle, R. W. (2015). Why is working memory capacity related to matrix reasoning tasks? *Memory & Cognition, 43*(3), 389–396. https://doi.org/10.3758/s13421-014-0473-3

Horn, W. (1983). *Leistungsprüfsystem: LPS [Performance Test System]*. Göttingen: Hogrefe.

Johnson, W., te Nijenhuis, J., & Bouchard, T. J., Jr. (2008). Still just 1 g: Consistent results from five test batteries. *Intelligence, 36*(1), 81–95. https://doi.org/10.1016/j.intell.2007.06.001

Kalra, P. B., Gabrieli, J. D., & Finn, A. S. (2019). Evidence of stable individual differences in implicit learning. *Cognition, 190*, 199–211. https://doi.org/10.1016/j.cognition.2019.05.007

Kan, K.-J., Kievit, R. A., Dolan, C., & van der Maas, H. (2011). On the interpretation of the CHC factor Gc. *Intelligence, 39*(5), 292–302. https://doi.org/10.1016/j.intell.2011.05.003

Kvist, A. V., & Gustafsson, J. E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's investment theory. *Intelligence, 36*(5), 422–436. https://doi.org/10.1016/j.intell.2007.08.004

Laurence, P. G., Mecca, T. P., Serpa, A., Martin, R., & Macedo, E. C. (2018). Eye movements and cognitive strategy in a fluid intelligence test: Item type analysis. *Frontiers in Psychology, 9*, 380. https://doi.org/10.3389/fpsyg.2018.00380

Loesche, P., Wiley, J., & Hasselhorn, M. (2015). How knowing the rules affects solving the Raven advanced progressive matrices test. *Intelligence, 48*, 58–75. https://doi.org/10.1016/j.intell.2014.10.004

Lozano, J. H., & Revuelta, J. (2020). Investigating operation-specific learning effects in the Raven's advanced progressive matrices: A linear logistic test modeling approach. *Intelligence, 82*, Article 101468. https://doi.org/10.1016/j.intell.2020.101468

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Nagy, G., Nagengast, B., Becker, M., Rose, N., & Frey, A. (2018). Item position effects in a reading comprehension test: An IRT study of individual differences and individual correlates. *Psychological Test and Assessment Modeling, 60*(2), 165–187.

Nagy, G., Ulitzsch, E., & Lindner, M. A. (2023). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning, 39*(3), 751–766. https://doi.org/10.1111/jcal.12719

Neubauer, A. C. (1990). Speed of information processing in the hick paradigm and response latencies in a psychometric intelligence test. *Personality and Individual Differences, 11*(2), 147–152. https://doi.org/10.1016/0191-8869(90)90007-E

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin, 142*(7), 758–799. https://doi.org/10.1037/bul0000046

Raven, J. C., & Raven, J. (2003). Raven progressive matrices. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 223–237). Springer. https://doi.org/10.1007/978-1-4615-0153-4_11.

Ren, X., Gong, Q., Chu, P., & Wang, T. (2017). Impulsivity is not related to the ability and position components of intelligence: A comment on Lozano (2015). *Personality and Individual Differences, 104*, 533–537. https://doi.org/10.1016/j.paid.2016.09.007

Ren, X., Schweizer, K., Wang, T., Chu, P., & Gong, Q. (2017). On the relationship between executive functions of working memory and components derived from fluid intelligence measures. *Acta Psychologica, 180*, 79–87. https://doi.org/10.1016/j.actpsy.2017.09.002

Ren, X., Schweizer, K., Wang, T., & Xu, F. (2015). The prediction of students' academic performance with fluid intelligence in giving special consideration to the contribution of learning. *Advances in Cognitive Psychology, 11*(3), 97–105. https://doi.org/10.5709/acp-0175-z

Ren, X., Wang, T., Altmeyer, M., & Schweizer, K. (2014). A learning-based account of fluid intelligence from the perspective of the position effect. *Learning and Individual Differences, 31*, 30–35. https://doi.org/10.1016/j.lindif.2014.01.002

Revelle, W. (2011). *An overview of the psych package*.

Ropovik, I. (2014). Do executive functions predict the ability to learn problem-solving principles? *Intelligence, 44*, 64–74. https://doi.org/10.1016/j.intell.2014.03.002

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Roth, M., & Herzberg, P. Y. (2008). Psychodiagnostik in der praxis: State of the art? *Klinische Diagnostik und Evaluation, 1*, 5–18.

Schweizer, K. (2013). A threshold-free approach to the study of the structure of binary data. *International Journal of Statistics and Probability, 2*(2), 67–76. https://doi.org/10.5539/ijsp.v2n2p67

Schweizer, K., Reiss, S., Schreiner, M., & Altmeyer, M. (2012). Validity improvement in two reasoning measures following the elimination of the position effect. *Journal of Individual Differences, 33*(1), 54–61. https://doi.org/10.1027/1614-0001/a000062

Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two-dimensional model of APM. *Psychology Science Quarterly, 51*(1), 47–64.

Schweizer, K., & Troche, S. (2018). Is the factor observed in investigations on the item-position effect actually the difficulty factor? *Educational and Psychological Measurement, 78*(1), 46–69. https://doi.org/10.1177/0013164416670711

Schweizer, K., Troche, S., Rammsayer, T., & Zeller, F. (2021). Inductive reasoning and its underlying structure: Support for difficulty and item position effects. *Advances in Cognitive Psychology, 17*(4), 274–283. https://doi.org/10.5709/acp-0336-5

Schweizer, K., Troche, S. J., & Rammsayer, T. H. (2011). On the special relationship between fluid and general intelligence: New evidence obtained by considering the position effect. *Personality and Individual Differences, 50*(8), 1249–1254. https://doi.org/10.1016/j.paid.2011.02.019

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science, 11*(6), 771–799. https://doi.org/10.1177/1745691616650647

Spilsbury, G., Stankov, L., & Roberts, R. (1990). The effect of a test's difficulty on its correlation with intelligence. *Personality and Individual Differences, 11*(10), 1069–1077. https://doi.org/10.1016/0191-8869(90)90135-E

Troche, S. J., Wagner, F. L., Schweizer, K., & Rammsayer, T. H. (2016). The structural validity of the culture fair test under consideration of the item-position effect. *European Journal of Psychological Assessment, 35*(2), 182–189. https://doi.org/10.1027/1015-5759/a000384

Van der Ven, A., & Ellis, J. L. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences, 29*(1), 45–64. https://doi.org/10.1016/S0191-8869(99)00177-4

Verguts, T., & De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement, 24*(2), 151–162. https://doi.org/10.1177/01466210022031589

Verguts, T., & De Boeck, P. (2002). The induction of solution rules in Raven's progressive matrices test. *European Journal of Cognitive Psychology, 14*(4), 521–547. https://doi.org/10.1080/09541440143000230

Vigneau, F., & Bors, D. A. (2008). The quest for item types based on information processing: An analysis of Raven's advanced progressive matrices, with a consideration of gender differences. *Intelligence, 36*(6), 702–710. https://doi.org/10.1016/j.intell.2008.04.004

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence, 34*(3), 261–272. https://doi.org/10.1016/j.intell.2005.11.003

von Gugelberg, H. M., Schweizer, K., & Troche, S. J. (2021). The dual mechanisms of cognitive control and their relation to reasoning and the item-position effect. *Acta Psychologica, 221*, Article 103448. https://doi.org/10.1016/j.actpsy.2021.103448

Wechsler, D. (2017). *Wechsler Intelligence Scale for Children* (Fifth Edition). Pearson.

Wiley, J., Jarosz, A. F., Cushen, P. J., & Colflesh, G. J. (2011). New rule use drives the relation between working memory capacity and Raven's advanced progressive matrices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 256–263. https://doi.org/10.1037/a0021613

Zeller, F., Krampen, D., Reiss, S., & Schweizer, K. (2017). Do adaptive representations of the item-position effect in APM improve model fit? A simulation study. *Educational and Psychological Measurement, 77*(5), 743–765. https://doi.org/10.1177/00131644166549

Zeller, F., Reiss, S., & Schweizer, K. (2017). Is the item-position effect in achievement measures induced by increasing item difficulty? *Structural Equation Modeling: A Multidisciplinary Journal, 24*(5), 745–754. https://doi.org/10.1080/10705511.2017.1306706