The Role of Rule Knowledge in Inductive Reasoning

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Psychologie und Sportwissenschaften
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Patrick Lösche
aus Lübbecke

Frankfurt am Main 2016
(D30)

Vom Fachbereich Psychologie und Sportwissenschaften der
Johann Wolfgang Goethe-Universität als Dissertation angenommen

Dekan:	Prof. Dr. Dr. Winfried Banzer
Erstgutachter:	Prof. Dr. Marcus Hasselhorn
Zweitgutachter:	Prof. Dr. Florian Schmiedek

Datum der Disputation: 22. Juni 2016

Man erblickt nur, was man schon weiß und versteht.

–Goethe (1819)

Danksagung

Zuallererst möchte ich Marcus Hasselhorn dafür danken, dass er mir überhaupt die Gelegenheit gegeben hat meine wissenschaftliche Neigung auszuleben und mich als Doktorand aufgenommen hat. Darüber hinaus hat er mich immer unterstützt und ermutigt. Danke dafür. Ich danke ferner allen Praktikanten und Hilfskräften die mich unterstützt haben: Christina Bader, Lothar Beck, Moritz Berghammer, Imke Bethcke, Elisabetta Cau, Celina Clément, Marlene Koch, Magdalena Kuhn, Nicoletta Lippert, Louisa Muschter, Marina Rempel. Ich danke all den Kolleginnen und Kollegen die mir mit kritischen Kommentaren und konstruktiven Anmerkungen weitergeholfen haben: Hanna Beißert, Janin Brandenburg, Judith Dirk, Frank Goldhammer, Allison Jaeger, Andrew Jarosz, Tilman Reinelt, Florian Schmiedek, Jennifer Wiley. Danke an Adam Chuderski und Miriam Vock für das Breitstellen von Aufgabenmaterial. Danke an Björn Rump und Filip Marinkovic für ein herausragendes Labormanagement. Danke auch an all die Schülerinnen und Schüler, Studentinnen und Studenten, und Lehrerinnen und Lehrer, die an den Studien beteiligt waren.

Inhalt

Zusammenfassung

Die vorliegende Dissertation befasst sich mit den kognitiven Prozessen die Intelligenz ausmachen. In diesem Zusammenhang wird der Frage nachgegangen, warum Aufgaben die das Arbeitsgedächtnis betreffen mit Aufgaben zur Intelligenzmessung korrelieren. Typische Aufgaben zur Intelligenzmessung sind zum Beispiel Matrix-Aufgaben. Diese lassen sich als Aufgaben zum induktiven Denken klassifizieren. Das bedeutet, dass dem Aufgabenmaterial eine versteckte Regel zugrunde liegt, die aus dem Aufgabenmaterials abgeleitet werden muss. Induktives Denken ist, vereinfacht gesagt, ein Rückschluss vom Einzelfall auf den allgemeinen Fall oder vergleichbare Fälle. Wenn also die versteckten Regeln anhand eines Teilproblems ausgemacht werden, so sollte man versuchen die Regeln auf andere Aspekte der Aufgabe anzuwenden. Eine zentrale Frage dieser Arbeit ist, ob aktuelle Modelle der Informationsverarbeitung, in denen das Arbeitsgedächtnis in der Regel eine zentrale Rolle spielt, erklären können wie Menschen solche Regeln herleiten. Die Arbeit ist in drei aufeinander aufbauende Teile aufgeteilt.

Der erste Teil gibt zunächst einen Überblick über die Intelligenzmessung bei Kindern und rezensiert den Beitrag kognitiver und nicht-kognitiver Variablen zur Vorhersage von Hochbegabung in einer Altersspanne die von der Geburt bis zur Einschulung reicht. Aus dem nicht-kognitiven Bereich stammen Konstrukte wie das Schlafverhalten, motivationale Faktoren wie Neugier und Interesse, und Einflussfaktoren aus dem sozialen Umfeld. Auch kognitive Variablen werden diskutiert, wie zum Beispiel frühe, außergewöhnliche Sprach-, Lese-, Schreib- und Rechenfähigkeiten, sowie Intelligenzquotienten die mit den gängigsten testpsychologischen Verfahren ermittelt werden. Außerdem werden Komponenten der Informationsverarbeitung wie Habituation und Arbeitsgedächtniskapazität als mögliche Prädiktoren diskutiert. Trotz der Berichte über mittlere Korrelationen ist die aktuelle Datenlage kritisch zu betrachten. Die meisten Verfahren weisen eine niedrige Vorhersagevalidität der Frühprognose von Hochbegabung auf. Das gilt insbesondere, je weiter man in das Säuglingsalter zurückgeht, wobei im vorschulischen Bereich Maße des Arbeitsgedächtnisses noch relativ vielversprechend sind. Es wird abschließend das dynamische Modell der Intelligenz angeführt, nach dem die mangelnde prognostische Validität und die Unzuverlässigkeit der kognitiven Vorhersageindikatoren auf die Annahme zurückzuführen sind, dass kognitive Prozesse in frühen Entwicklungsstadien

noch weitgehend unabhängig voneinander sind. Erst im Laufe des Lebens werden diese immer häufiger miteinander verknüpft und korreliert (Mutualismus). Das bedeutet, es findet eine zunehmende Integration kognitiver Prozesse statt und daraus resultiert der g-Faktor.

Doch das Arbeitsgedächtnis spielt weiterhin eine zentrale Rolle in Bezug auf Intelligenz, wenn nicht im Säuglingsalter, dann zumindest ab dem Vorschulalter und besonders im Erwachsenenalter. Arbeitsgedächtniskapazität stellt das Ausmaß an Fähigkeit dar, Informationen simultan zu speichern und zu verarbeiten, ohne die Notwendigkeit auf Vorwissen zurückzugreifen. Es kann damit als eng umschriebener kognitiver Prozess aufgefasst werden. Vorherige Forschungsarbeiten haben bereits deutlich gezeigt, dass Arbeitsgedächtnis und Intelligenz korrelativ stark zusammenhängen. Das ist durchaus überraschend, denn die Aufgaben und Tests mit denen die Konstrukte jeweils erfasst werden können sich oberflächlich stark unterscheiden. Um fluide Intelligenz valide erfassen zu können, sind Aufgaben zum induktiven Denken, wie zum Beispiel Matrix-Aufgaben, sehr beliebt. Wie bereits erwähnt, ist eine zentrale Charakteristik dieser Aufgaben, dass versteckte Regeln entdeckt werden sollen. Es liegt also unvollständige Information vor. Auf der anderen Seite liefern typische Aufgaben zu Messung der Arbeitsgedächtniskapazität, wie zum Beispiel komplexe Spannenaufgaben, vollständige Information. Das bedeutet, bei Arbeitsgedächtnisaufgaben ist die Aufgabenleistung allein durch das Kapazitätslimit begrenzt und die Aufgabenschwierigkeit rührt daher, dass das zu merkende Material im Umfang zunimmt während gleichzeitige Ablenkung einwirkt. Bei Arbeitsgedächtnisaufgaben muss also nicht hergeleitet oder entdeckt werden.

Der zweite Teil der vorliegenden Arbeit widmet sich daher dem Einfluss der Prozesse Zielmanagement und Regel-Induktion auf den Zusammenhang zwischen Arbeitsgedächtnis und Problemlösung der Matrix-Aufgaben. Basierend auf vorherigen Forschungsarbeiten war die Ausgangsvermutung, dass Zielmanagement neben Regelinduktion die zweite Subkomponente des Löseprozesses bei Raven‘s Advanced Progressive Matrices darstellt. Zielmanagement wurde bereits zuvor mit dem Arbeitsgedächtnis in Verbindung gebracht jedoch war die Befundlage hinsichtlich der Regel-Induktion unklar.

Daher sollte die Hypothese ob Regel-Induktion unabhängig vom Arbeitsgedächtnis ist, in einem kritischen Experiment überprüft werden (N = 644, mittleres Alter = 12 Jahre). Bei Neutralisierung der Notwendigkeit für Regelinduktion, indem den Probanden die Regeln im Voraus erklärt wurden, konnte man einen erhöhten Zusammenhang zwischen der Arbeitsgedächtniskapazität und den Ergebnissen im Raven-Test erkennen. Das deutet in der Tat darauf hin, dass Regel-Induktion nicht vom Arbeitsgedächtnis abhängig ist, zumindest nicht im gleichen Maße wie Zielmanagement. Darüber hinaus zeigte sich, dass die Kenntnis der Regeln den Problemlöseprozess beeinflusst und die Aufgabenleistung insgesamt deutlich erhöht.

Dieser Befund wurde im Wesentlichen in zwei weiteren Experimenten bestätigt (jeweils *N* = 366 und *N* = 393, mittleres Alter = 12 Jahre). Jedoch sollte in diesen weiteren Experimenten auch eine Komplementärhypothese überprüft werden. Demnach sollte, wenn bei bekannten Regeln keine Regel-Induktion mehr notwendig wäre, nicht nur die Korrelation mit Arbeitsgedächtnisaufgaben steigen, sondern auch umgekehrt die Korrelation mit Aufgaben die für die Regel-Induktion relevante Prozesse erfassen fallen. Um die kognitiven Prozesse der Regel-Induktion messbar zu machen wurde in einem Experiment die „Brixton Rule Anticipation“ Aufgabe eingesetzt. Bei dieser Aufgabe geht es darum das Bewegungsmuster eines Punktes vorherzusagen, wobei sich dieser nach einer bestimmten Regel innerhalb einer räumlichen Anordnung bewegt. Diese Aufgabe wurde zuvor bereits in Neuropsychologischen Experimenten eingesetzt und man vermutet dass die Prozesse der Regel-Entdeckung und Regel-Anwendung eine zentrale Rolle spielen. In einem weiteren Experiment wurden typischen Aufgaben aus der Kreativitätsforschung eingesetzt. Bei diesen Aufgaben geht es darum innerhalb einer vorgegebenen Zeitspanne möglichst viele Ideen oder Begriffe zu einem Schlagwort zu generieren und aufzuschreiben. Es wurden sowohl die Anzahl der generierten Ideen gewertet (Flüssigkeit) als auch deren Qualität (Kreativität). Jedoch zeigten sich bei keiner der beiden Aufgaben die vorhergesagten Ergebnismuster. Die Korrelationen waren zwischen den Experimentellen Bedingungen nicht signifikant verschieden. Eine mögliche Ursache für diesen Befund war, dass die Aufgaben möglicherweise keine reinen Maße der Regel-Induktion darstellen. Die Brixton Aufgabe schien eine starke Arbeitsgedächtnisanforderung zu haben und die Maße zur Kreativität hatten eine bedenkliche Reliabilität.

In einem vierten Experiment wurde das gleiche Paradigma wie in den ersten drei Experimenten eingesetzt jedoch wurden nun zusätzlich Augenbewegungen erfasst (*N* = 47, mittleres Alter = 19 Jahre). Der Hintergrund war, dass bereits einige Vorarbeiten in diesem Bereich gezeigt haben, dass Augenbewegungsmuster auf gewisse Lösungsstrategien hindeuten und weiterhin, dass diese auch von der

Arbeitsgedächtniskapazität der jeweiligen Person abhängig sind. Es stand daher die Hypothese im Raum, ob nicht das Wissen der Regeln in der einen experimentellen Bedingung die Anwendung von Anspruchsvollen Lösungsstrategien gefördert hat. Tatsächlich lag die Analyse der Augenbewegungen genau diesen Schluss nahe. Diese ergab, dass Probanden in der Bedingung mit bekannten Regeln eine Strategie anwenden bei der eine potenzielle Antwort im visuellen Arbeitsgedächtnis erstellt wird um diese dann mit den vorgegebenen Antwortalternativen abzugleichen (*engl.* constructive matching). Damit eröffnete sich eine alternative Erklärung zu den Befunden aus den ersten drei Experimenten. Die Korrelation mit Maßen der Arbeitsgedächtniskapazität hätte sich demnach erhöht, weil anspruchsvollere Strategien zum Einsatz gekommen sind, die zwar effektiver sind, gleichzeitig jedoch eine zusätzliche Beanspruchung für das Arbeitsgedächtnis darstellen.

Der dritte und letzte Teil der vorliegenden Arbeit widmete sich deshalb erneut dieser Fragestellung und stellt die Ergebnisse aus zwei weiteren Experimenten dar (jeweils *N* = 50 und *N* = 109 aus einer studentischen Population). Erneut kamen Matrix-Aufgaben zum Einsatz und es wurden Testleistungen, Augenbewegungen und Reaktionszeiten erhoben, um den Einfluss von Regelwissen zu erfassen. Es zeigte sich erneut die Anwendung einer effektiveren Lösungsstrategie in der Experimentalbedingung. Anhand der Eye-Tracking Messung wurde gezeigt, dass Probanden mit Regelwissen über einen längeren Zeitraum das problemrelevante Areal der Matrix-Aufgabe fixieren, und eine niedrigere Frequenz an Sakkaden zwischen diesem Areal und den Antwortalternativen aufwiesen. Weitere Einflussvariablen auf die Lösestrategie stellen Schwierigkeit der Aufgabe und Fähigkeiten des Probanden dar. Diese weisen einen differenziellen Einfluss auf zwei Subgruppen von Indikatoren der Augenbewegungsmessung auf, die in Relation zu den Reaktionszeiten gesetzt wurden um ein besseres Verständnis dieser Variablen zu erzielen. Es wird vermutet, dass Variablen wie Augenbewegungen und Reaktionszeiten das Ausmaß des Entstehens von mentalen Modellen während des logischen Denkens widerspiegeln. Unter der Annahme dass die Komplexität von Mentalen Modellen mit einer gewissen Belastung für das Arbeitsgedächtnis einhergeht, lassen sich auch vorherige Ergebnisse mit dieser Hypothese in Einklang bringen.

Abschließend werden die grundlegenden kognitiven Prozesse des induktiven Denkens diskutiert. Es wird eine Theorie ausgearbeitet die affektive Reaktionen und motivationale Prozesse berücksichtigt und induktives Denken im Wesentlichen als

Resultat des Zusammenspiels von Langzeitgedächtnis und Arbeitsgedächtnis betrachtet. Zu guter Letzt wird nochmal ausgeführt warum es wichtig ist, Intelligenz nicht nur als psychometrisches Konstrukt und Kraft mit unbekanntem Ursprung aufzufassen, sondern sich stattdessen auch in der differenziellen Psychologie mit eng umschriebenen kognitiven Prozessen zu befassen.

Prologue

What is intelligence? Many students of psychology learn that intelligence is what the tests test. But what does that mean? This frequently cited definition was originally written by Boring (1923) as part of a commentary on the then-current state of research in psychological measurement. He was concerned with the validity of measures of intelligence and discussed the implications for the assessment of cognitive development and heredity. Boring noted that contemporary scientists have made good progress in developing measures of cognitive ability that could reliably discriminate between those who did well and those who did not. But he also noted that there was no single measure of intelligence available, instead there were batteries of tests, and each test could be quite different from the other. Today, research on the measurement of intelligence has unveiled some of the cognitive processes involved in reasoning and problem solving. But still, many common measures of intelligence are constructed as batteries of various tests, grounded in the psychometric theory of intelligence.

This theory was pioneered by Spearman (1904). His seminal work on the objective measurement of general intelligence made evident, for the first time, that many tests have a common underlying factor. Spearman was unhappy with the state of experimental psychology and reviewed a host of findings that seemed largely incompatible with each other. Many of his contemporary psychological scientists were trying to assess intelligence and the approaches were as numerous as they were diverse. These included, for example, the assessment of reaction time, memory, attention, sight and hearing, creativity, weight discrimination, moral sentiments, muscular force, and even swiftness in fencing. Sometimes, the tests were compared to teacher ratings of their pupils' aptitudes, but oftentimes they were simply compared among each other. In the eye of Spearman, nothing conclusive came of this line of research, as some researchers reported correlations that were contradicted by others. He was particularly disappointed with the lack of validity of laboratory research for real world practical intelligence, and went on to introduce the concepts of measurement error and confounds. Not only did the consideration of these concepts inspire classical test theory, but also lead to the invention of factor analysis.

It is important to point out that Spearman (1904) confined himself to the measurement of simple perceptual discrimination of weight, sound, and sight. The sight-task required participants to judge the difference of the brightness of grayscale printed cards. The dependent variable was the minimal difference in brightness that could reliably be detected. The weight-task worked exactly the same, except that weights were compared (controlled for their size). Finally, the sound-task required the discrimination of tone pitch in the same way. He named these measurements “the discriminations” and compared them with “the intelligences”. But Spearman did not directly measure intelligence with tests on his subjects, which were all school children from various schools in his neighborhood. Instead, Spearman asked teachers to classify the students in high, low, and average brightness. He also asked the headmaster’s wife, and two of the oldest children of a class to do the same classifications in order to obtain multiple measurements. Spearman collected these data on five different occasions, amounting to a combined sample size of 123 boys and girls. Although the methods to measure or estimate intelligence might appear crude in light of today’s advances in psychological measurement, the strength of this work lies predominantly in the theoretical thoughts and statistical analyses. Spearman noted that all measurements of discrimination were positively correlated and could be statistically accounted for by a general discrimination factor. The same was true for all ratings of intelligence, which could be accounted for by a general intelligence factor. Furthermore, both factors were almost perfectly correlated. The general factor of intelligence was born: the g-factor.

From that point on, intelligence was mostly equated with the g-factor and was seen as some mysterious force that accounts for performance in all sorts of tasks that require thought (Spearman, 1923). The phenomenon of positive correlations among all sorts of tests is known as the positive manifold and reveals one fundamental characteristic of human intelligence: Omnipresence. That is, if someone is good at one particular task (e.g., math), chances are that they are also good at any other task (e.g., arts). The positive manifold implies that all tasks are influenced by one central force, and if that force is strong, then it has an impact on everything. This means that there can hardly be any trade-off between different abilities, however Spearman did point out that “at any rate to assume anything like a *universal* correspondence of this kind – so that, for example, the man with greater power of imagination for chess must necessarily have it also for music – is a palpable fallacy” (Spearman, 1923, p. 4). This notion gave rise to the theory of two factors, proposing that every task is partly determined by some specific source of variance. Yet, the major force is always the g-factor or whatever is responsible for it.

Even today, more than a century after Spearman's revelation, there is no definitive answer as to the cause of the g-factor. Some researchers pursued the idea that it is grounded in elementary cognitive speed (e.g., Jensen, 1993). More recently, working memory seems a promising candidate (see Ackerman, Beier, & Boyle, 2005). Yet another theory proposes that there may be no single source for the g-factor and views it as an epiphenomenon (van der Maas et al., 2006). The current dissertation is an attempt to identify at least some of the cognitive processes that are relevant for intelligence. In order to find an answer to the question what intelligence is, I will ask: What are the cognitive mechanisms and processes that are relevant in intelligence tests? What does a human brain (or any brain for that matter) have to accomplish to perform well on such tests? There is little dispute about regarding someone who performs well on IQ tests as being intelligent. In this sense: Intelligence is what the tests test. What are they testing?

Part 1: The Prospects and the Limits to the Prediction of Giftedness

Publication Note

The following part was published in German language as: Beißert, H., Hasselhorn, M., & Lösche, P. (2014). Möglichkeiten und Grenzen der Frühprognose von Hochbegabung. In M. Stamm (Ed.), *Handbuch Talententwicklung* (pp. 415-425). Bern, Switzerland: Hans Huber.

Abstract

This chapter is intended to provide an overview of the current state of research in assessment and prediction of talent and giftedness. The focus is on the age range from infancy to preschool. Non-cognitive variables, such as sleeping habits, social background, and motivational factors, are discussed. Cognitive indicators, like early information processing, preschool intelligence, and exceptional preschool precociousness in literacy and math are discussed as well. In conclusion, the possibility to predict giftedness appears to be unreliable for the most part of this age range. One possible explanation is discussed by considering a dynamic model of general intelligence that provides a theoretical account for limited reliability of intelligence assessment during early development.

Einleitung

Außergewöhnliche Begabungen üben seit jeher Faszination aus. In der wissenschaftlichen wie öffentlichen Diskussion steht dabei oftmals die Frage im Vordergrund, was denn unter Hochbegabung zu verstehen sei. Anfang der 1970er Jahre veröffentlichte die amerikanische Regierung einen Bericht über die Bildungsmöglichkeiten und Förderinitiativen für Hochbegabte. Da dieser unter der Federführung des damaligen Bildungsbeauftragten Sidney Marland entstand, wird er oftmals als Marland-Bericht bezeichnet. Hochbegabung wurde im Marland-Bericht als herausragendes Verhaltenspotential in einem oder mehreren der folgenden Bereiche definiert: allgemeine intellektuelle Fähigkeit, spezifische schulische Hochleistungen, kreatives oder produktives Denken, Führungsqualität, bildnerische und darstellende Künste sowie psychomotorische Fähigkeiten (Marland, 1971). Dadurch wird ein sehr breites Spektrum von Hochbegabung aufgespannt. Wissenschaftliche Bemühungen um eine präzise Definition von Hochbegabung standen immer wieder vor dem Problem, dass Personen mit individuellen Spitzenwerten in einem der Bereiche, oftmals in den meisten anderen Bereichen keine Spitzenwerte erzielen. In der langen Tradition psychometrischer Intelligenztests hat sich aber immer wieder gezeigt, dass die Intelligenztestleistungen zu den besten Prädiktoren für ganz unterschiedliche kulturelle Leistungen zählen. Daher reserviert man heute in der Regel den Begriff Hochbegabung für das Phänomen besonders ausgeprägter allgemeiner intellektueller Fähigkeiten (Rost, 2009a, 2009b).

Die festgestellte Hochbegabung einer Person (meist operationalisiert über einen IQ ≥ 130) hat also hohen prognostischen Wert für spätere Leistungen. Wann und wie aber lässt sich im frühen Kindesalter prognostizieren, ob ein Kind im Schulalter oder später das Hochbegabungs-Kriterium erfüllt? Diese Frage steht im Fokus des vorliegenden Beitrags.

Alle gängigen Theorien der allgemeinen Intelligenz gehen davon aus, dass die intellektuelle Grundausstattung des Menschen zu einem beträchtlichen Teil biogenetisch determiniert ist (Neisser et al., 1996). Von Geburt an sollte daher feststehen, in welcher Bandbreite der messbare IQ später einmal liegen wird. Diese Annahme nährt die Hoffnung, dass man schon im frühen Kindesalter Verhaltensmarker finden kann, mit deren Hilfe vergleichsweise sicher vorhersagbar ist, ob ein Kind im Schulalter einen IQ von 130 oder mehr aufweisen wird.

Der Beitrag gibt einen Überblick über die Ansätze und empirischen Ergebnisse bisheriger Bemühungen der Frühprognose von Hochbegabung. Insgesamt zeigt sich dabei, dass die prognostischen Validitäten selbst der besten identifizierten Verhaltensmarker eher bescheiden ausfallen, so dass Einzelfallprognosen etwa vom Kleinkindalter auf eine Hochbegabung im Schulalter nicht zuverlässig möglich sind. Das dynamische Modell der allgemeinen Intelligenz von van der Maas et al. (2006) sagt im Unterschied zu klassischen g-Faktor-Modellen der Intelligenz genau dies auch voraus. Der vorliegende Beitrag endet daher mit einer Skizze dieses aktuellen Ansatzes.

Empirische Versuche der Frühprognose

In der Literatur finden sich eine Reihe verschiedener Verfahren bzw. Indikatoren, die herangezogen wurden, um zu einem möglichst frühen Zeitpunkt im Leben eines Kindes eine später nachweisbare Hochbegabung festzustellen. Im Folgenden werden die gängigsten dieser Indikatoren vorgestellt und diskutiert.

Schlafverhalten

In der Praxis ist die Überzeugung weit verbreitet, dass hochbegabte Kinder ein verringertes Schlafbedürfnis aufweisen oder gar unter Schlafproblemen leiden (Freeman, 1979; Stapf, 2010). Dies spiegelt sich auch in einigen Checklisten zur Erkennung Hochbegabter wieder (Graue, 1985) . Die Annahme, dass kindliches Schlafverhalten als Identifikationsmerkmal von Hochbegabung geeignet ist, wird durch die empirische Forschungslage allerdings nur wenig gestützt. Es finden sich zwar Studien, in denen die hochbegabten Kinder kürzer schliefen als die Referenzgruppe (Stapf & Stapf, 1988), aber auch Studien, in denen Umgekehrtes der Fall war (Jung, Molfese, Beswick, Jacobi-Vessels, & Molnar, 2009; Terman, 1925), oder in denen sich überhaupt keine Unterschiede im Schlafverhalten feststellen ließen (Freeman, 1979). Zudem ist die Forschung in diesem Bereich nur begrenzt aussagefähig, denn bei den meisten Studien handelt es sich um Retrospektivbefragungen und/oder Studien mit relativ kleinen Stichprobengrößen, die auch häufig erst mit Kindern ab dem Schulalter durchgeführt wurden (Perleth, Schatz, & Mönks, 2000). Prospektive Längsschnittuntersuchungen, in denen das Schlafverhalten im frühen Kindesalter als Prädiktor für eine Hochbegabung in späteren Altersstufen herangezogen wird, fehlen fast vollkommen.

Bei einer Stichprobe von mehr als 17.000 Kindern fand Pollock (1992) keine nennenswerten Zusammenhänge zwischen Schlafproblemen in früher Kindheit und den intellektuellen Fähigkeiten im Alter von zehn Jahren. Allerdings wurden in dieser Studie nicht speziell hochbegabte Kinder untersucht und nicht gezielt nach der Schlafdauer gefragt. In einer Längsschnittstudie von Jung et al. (2009) wiesen Kinder mit täglichen Schlafzeiten von über 8 Stunden zu allen Messzeitpunkten (3, 4 und 5 Jahre) höhere Werte in den kognitiven Maßen auf. Es findet sich dort jedoch kein Zusammenhang zwischen der Schlafdauer und einer beschleunigten kognitiven Entwicklung, was zu erwarten gewesen wäre, wenn die Schlafdauer eine kausale Ursache für das Tempo der kognitiven Entwicklung wäre. Zusammenfassend ist zu sagen, dass vor dem Hintergrund der aktuellen Befundlage das kindliche Schlafverhalten nicht zuverlässig als früher Indikator von Hochbegabung herangezogen werden kann (Perleth et al., 2000).

Außergewöhnliche Leistungen

Frühe und besondere Leistungen in Bereichen wie Sprechen, Lesen, Schreiben oder Rechnen werden oft als Hinweis auf eine Hochbegabung interpretiert (Stöger, Schirner, & Ziegler, 2008). Betrachtet man allerdings die Forschungslage, so zeigt sich, dass dies nicht unbedingt gerechtfertigt ist. Sowohl beim frühen Lesen als auch beim frühen Rechnen ist die Befundlage nicht eindeutig. Jackson, Donaldson, and Cleland (1988) konnten zwar schwache Zusammenhänge zwischen frühen Lesekompetenzen und Intelligenz finden, allerdings kommen sie zu dem Schluss, dass eine besonders hohe Intelligenz keine notwendige Bedingung für den Erwerb früher Lesekompetenzen ist, sondern dass frühes Lesen von einer Reihe anderer Aspekte abhängt.

Stamm (2004) demonstrierte in einer Längsschnittstudie, dass die Gruppe der Frühleser und Frührechner auch acht Jahre nach der Einschulung noch einen signifikanten Vorteil in Intelligenz und Schulleistungen aufweisen. Trotz dieses Zusammenhangs wird dort auch deutlich, dass hohe Intelligenz keine zwingende Voraussetzung ist, um früh Lesen oder Rechnen zu lernen. Es finden sich viele später nicht hochbegabte Kinder unter den frühen Lesern und Rechnern und umgekehrt auch viele Hochbegabte, die nicht früh mit dem Lesen oder Rechnen begonnen haben. Vielmehr scheint die Eigeninitiative beim vorschulischen Kompetenzerwerb ein wichtiger Faktor zu sein. So zeigt sich zum Beispiel bei Schülern, deren vorschulischer Kompetenzerwerb auf Instruktionen der Eltern zurückgeht, ein erhöhtes Risiko ihren Leistungsvorsprung später zu verlieren. Stamm (2004) stellt fest, dass „die elterliche Anleitung zum Lesen- oder Rechnenlernen lediglich eine vergleichsweise unbedeutende Rolle spielt“ (S. 405). Ferner berichtet Stamm (2004), dass Kinder aus der Mittel- und Oberschicht eher früh lesen und rechnen, jedoch auch 20% der Kinder aus bildungsfernem Milieu zu den Frühlesern und Frührechnern gezählt werden können, somit also nicht auf ein intellektuell-responsives Elternhaus angewiesen sind. Sie vermutet, dass sich Kinder mit hoher Eigeninitiative auch bei intellektuell wenig anregenden Umgebungen die notwendigen Umweltbedingungen selbst schaffen.

Im Bereich des frühen Schreibens gibt es wenig Forschung, da frühe Schreibfertigkeiten durch feinmotorische Fähigkeiten beschränkt werden, die sich gewöhnlich nicht vor dem fünften Lebensjahr entwickeln (Rudolf, 1980). Durch diese asynchrone Entwicklung von Lesen und Schreiben werden Kinder, obwohl sie bereits lesen können, möglicherweise durch mangelnde Feinmotorik darin behindert, Sprache in Schriftform darzustellen (Terassier, 1985).

Etwas günstiger fallen die Befunde zur frühen Sprachentwicklung für die Prognose von späterer Hochbegabung aus. So finden sich Zusammenhänge zwischen früher und schneller Sprachentwicklung und verschiedenen Intelligenzmaßen (Robinson, Dale, & Landesman, 1990). In der in diesem Zusammenhang am häufigsten zitierten Belegstudie von Cameron, Livson, and Bayley (1967) zeigte sich allerdings ein Interaktionseffekt: der Zeitpunkt der ersten kindlichen Wortproduktion erwies sich nur bei Mädchen (nicht bei Jungen) als moderater Prädiktor für den IQ im jungen Erwachsenenalter. Andere Studien berichten lediglich aggregierte Effekte, die sich nicht ohne Weiteres auf die Individualdiagnostik übertragen lassen (Shapiro et al., 1989). Somit sind auch die prognostischen Möglichkeiten der frühen Sprachentwicklung sehr beschränkt. Insgesamt lässt sich also festhalten, dass außergewöhnliche Leistungen in der frühen Kindheit nicht zuverlässig auf eine spätere Hochbegabung hindeuten.

Motivationale Aspekte

In vielen Modellen zur Hochbegabung ist Motivation eines der relevanten Kriterien, um hohes intellektuelles Potenzial in außergewöhnliche Leistungen umzuwandeln (Gagné, 1993; Heller, Perleth, & Hany, 1994). Da sich aber motivationale Dispositionen oftmals erst im Laufe der Grundschulzeit entwickeln, gibt es kaum Untersuchungen, die die Motivation vor dem Schulalter untersuchen (Perleth et al., 2000). In der frühen Kindheit scheinen aber Neugier und Interesse relevante motivationale Aspekte für die Identifikation von Hochbegabung zu sein (Winner, 2004). Neugier, die sich in der frühen Kindheit vor allem im kindlichen Explorationsverhalten manifestiert, stellt einen motivationalen Zustand dar, der Kinder dazu antreibt, sich neuen Reizen auszusetzen und Informationen aufzunehmen (Berg & Sternberg, 1985; Lehwald, 1991). Das soziale Umfeld reagiert auf die Neugier des Kindes mit der Bereitstellung besonderer Lerngelegenheiten und treibt somit die Entwicklung der kognitiven Fähigkeiten voran (Stöger et al., 2008). So zeigt sich, dass das Interesse an Neuem in der frühen Kindheit prädiktiv für die spätere Intelligenz ist (Berg & Sternberg, 1985).

Allerdings ist nun auch schon ein Problem angesprochen, das mit dem Heranziehen von Neugier als Identifikationsmerkmal von Hochbegabung verbunden ist: Die Entwicklung von motivationalen Ausprägungen in der Kindheit ist eng an das familiäre Umfeld gebunden und somit nicht von der Umwelt des Kindes zu trennen.

Auszugehen ist hier von reaktiven Anlage-Umwelt-Interaktionen, da die Umwelt auf das Kind reagiert und es durch diese Reaktionen weiter fördert (Perleth, 2000).

Soziales Umfeld

Von Anzeichen einer Hochbegabung im frühen Kindesalter lässt sich nicht ohne Weiteres auf eine Hochbegabung in späteren Lebensphasen schließen (Stöger et al., 2008). Lewis and Michalson (1985) sehen eine Erklärung dafür im Einfluss des sozialen Umfelds, vor allem des familiären Hintergrunds, auf die Begabungsentwicklung. Neben der genetischen Mitgift schaffen Eltern auch Entwicklungsmöglichkeiten (zum Beispiel in Form von Lernumwelten) und leisten somit aktiv einen Beitrag an der Begabungsentwicklung ihrer Kinder (Lewis & Michalson, 1985; Stöger et al., 2008). Es finden sich Hinweise, dass dieses Bereitstellen von Lerngelegenheiten dem Einfluss sozialer Schichtungsfaktoren unterliegt (Büchner & Krüger, 1996). So scheinen Eltern aus intellektuellen Milieus besser in der Lage zu sein, ihren Kindern Lernumwelten zu bieten, die für die Entwicklung der kognitiven Fähigkeiten förderlich sind. Interessant ist dabei auch, dass der Bildungshintergrund der Familie und die Lernumwelt des Kindes hoch mit dem sozioökonomischen Hintergrund der Familie korrelieren (Stöger et al., 2008). Dies könnte ein Grund dafür sein, warum hochbegabte Kinder überzufällig häufig aus mittleren bis oberen sozialen Schichten stammen (Carman & Taylor, 2010; Rost, 1993).

Das soziale Umfeld darf also in seinem Einfluss auf die Genese von Begabung nicht unterschätzt werden. In einer Studie von Rubin and Balow (1979) erwies sich der sozioökonomische Hintergrund der Eltern sogar als besserer Prädiktor für Hochbegabung als verschiedene frühe Intelligenzmaße.

Intelligenzmessung anhand psychometrischer Verfahren

Intelligenztests spielen nach wie vor eine bedeutende Rolle in der Intelligenz- und Hochbegabungsdiagnostik. Es gibt eine Reihe von Tests, die sich schon im frühen Kindesalter anwenden lassen. Die Stanford-Binet Intelligenz-Skala (Thorndike, Hagen, & Sattler, 1986), die Wechsler Preschool and Primary Scale of Intelligence (WPPSI-III) (Wechsler & Petermann, 2009) und die Kaufman Assessment Battery for Children (K-ABC) (Kaufman, Kaufman, Melchers, & Preuß, 2009) sind z.B. häufig eingesetzte Verfahren, die bereits ab einem Alter von zwei bis drei Jahren angewendet werden können (Helmsen, Lehmkuhl, & Petermann, 2009). Eine sehr umfassende Bewertung verschiedener Intelligenztests für das frühe Kindesalter findet sich bei Perleth et al.

(2000). Eine Sichtung der Literatur ergibt allerdings, dass die größten Schwächen von Intelligenztests im frühen Kindesalter in deren mangelnder Reliabilität und geringen Entwicklungsstabilität liegen (Sattler, 1988). Es zeigt sich immer wieder, dass Ergebnisse aus Intelligenztests im (frühen) Kindesalter nur gering bis maximal moderat mit solchen aus dem Erwachsenenalter korrelieren (Schneider, Bullock, & Sodian, 1998; Shapiro et al., 1989).

Intelligenzmessungen im frühen Kindesalter scheinen also auch nicht dazu geeignet zu sein, zuverlässig eine spätere Hochbegabung vorherzusagen (Shapiro et al., 1989; Stöger et al., 2008). Die Stabilität von Intelligenzmessungen nimmt erst im Laufe der Schulzeit deutlich zu (Perleth et al., 2000). Prinzipiell gilt dabei, je älter das Kind bei der ersten Testung ist und je geringer das Retestintervall, desto stabiler ist der IQ (vgl. Sattler, 1988, S. 73).

Komponenten der Informationsverarbeitung

Es hat seit den 1970er Jahren immer wieder Ansätze gegeben, die Hochbegabung auf eine besonders effektive und effiziente Informationsverarbeitung zurückführen (Sternberg, 1985). In der frühen Kindheit gemessen, stellen Maße der Informationsverarbeitung einen moderaten Prädiktor für IQ-Werte im mittleren Kindesalter dar (Rose, Feldman, Jankowski, & Van Rossem, 2012). Typische Aspekte, die in diesem Rahmen untersucht werden, sind Habituation, Gedächtnis, Verarbeitungskapazität und Informationsverarbeitungsstrategien (Stöger et al., 2008).

Das Habituationsparadigma

Besonders Rekognition und Habituation scheinen auf den ersten Blick für die Prognose von Hochbegabung geeignet zu sein. Diese Verfahren nutzen aus, dass Säuglinge und Kleinkinder in der Regel mehr Aufmerksamkeit auf neue Reize lenken, diese also gegenüber vertrauten Reizen länger und häufiger anschauen. Die Häufigkeit oder Dauer der Darbietung eines neuen Reizes (z.B. ein Spielzeug), bis er gegenüber der erstmaligen Darbietung nur noch halb so viel Interesse des Kindes auf sich zieht, wird als Habituation oder Habituierung bezeichnet und gilt als Indikator dafür, dass das Kind den Reiz im Gedächtnis gespeichert hat. Dementsprechend kann man schlussfolgern, dass sich ein Kind an einen Reiz erinnert (Rekognition), wenn es diesen nach einer gewissen Pause wieder vorgelegt bekommt und wenig Interesse daran zeigt (McCall & Carriger, 1993).

In einem umfangreichen Review kommen McCall and Carriger (1993) zu dem Schluss, dass derartige Maße der Informationsverarbeitung im Säuglingsalter zu einem gewissen Grad den IQ der späten Kindheit oder sogar den IQ des jungen Erwachsenenalters vorhersagen können. Basierend auf einer Metaanalyse schätzen sie die mittlere Korrelation etwa auf .36. Dieser Wert wurde von Fagan, Holland, and Wheeler (2007) in einer neueren Längsschnittstudie bestätigt. Wenn man die Effekte des sozioökonomischen Status auspartialisiert und für die Reliabilität der Maße minderungskorrigiert, so zeigen sich Korrelationen von bis zu .59 mit schulischen Leistungen oder der Intelligenz im jungen Erwachsenenalter. Auch Rose et al. (2012) kommen auf der Grundlage einer Längsschnittstudie zu ähnlich hohen prädiktiven bzw. prognostischen Validitäten. Verschiedene Maße für die Informationsverarbeitung im Säuglings- und Kleinkindalter korrelieren (unkorrigiert) bis zu .26 mit dem Wechsler-IQ im Alter von 11 Jahren.

Allerdings muss dabei beachtet werden, dass bisher keine Studien speziell zur Prognose von Hochbegabung durchgeführt wurden. So kann nicht ausgeschlossen werden, dass die Korrelationen überwiegend durch Varianz zustande kommt, die durch besonders schwache Kinder erzeugt wird (McCall & Carriger, 1993). Ein weiteres nicht zu vernachlässigendes Problem von Habituierungs- und Rekognitionsmaßen sind deren geringe Retestreliabilitäten (Slater, 1997). Zwar zeigen die minderungskorrigierten Korrelationen deutlich höhere Werte, aber diese sind eher entwicklungspsychologisch interessant, weil sie auf eine gewisse Kontinuität in der Entwicklung der kognitiven Prozesse (differentielle Entwicklungsstabilität) schließen lassen. Sollten die Tests jedoch für die Einzelfalldiagnostik verwendet werden, so wird die prognostische Validität durch die schwachen Reliabilitäten stark eingeschränkt. Daraus lässt sich schließen, dass der Nutzen von Rekognitions- und Habituierungsmaßen zur Identifikation und Auswahl von Hochbegabten sehr beschränkt ist. Will man sie nutzen, so ist zu beachten, dass im Alter zwischen zwei und acht Monaten die Langzeitprognosen besonders gut ausfallen (McCall & Carriger, 1993).

Arbeitsgedächtniskapazität

Arbeitsgedächtniskapazität wird typischerweise mit Aufgaben erfasst, die simultane Speicherung und Verarbeitung von Informationen verlangen. Die Bewältigung dieser Aufgaben erfordert keinerlei Vorwissen und wird als relativ pures Maß der Informationsverarbeitungskapazität verstanden. Im Gegensatz dazu erfassen globale Intelligenztests eine Vielzahl an kognitiven Prozessen. Alloway and Alloway (2010) berichten, dass die Arbeitsgedächtniskapazität im Alter von 4 bis 5 Jahren die sechs Jahre später beobachtbaren schulischen Leistungen im Alter von 10 bis 11 Jahren vorhersagen kann. Die Korrelationen liegen im Bereich von .33 bis .45 und sind ähnlich hoch wie die für zwei Subtests aus dem Wechsler Vorschul-Intelligenztest. Die Ergebnisse der von den Autoren durchgeführten Regressionsanalysen legen weiterhin den Schluss nahe, dass Arbeitsgedächtniskapazität eigene Varianzanteile der Schulleistung im Vergleich zum nonverbalen IQ vorhersagen kann. Der Ansatz ist also recht vielversprechend, jedoch dadurch eingeschränkt, dass noch wenige Studien zum Vorschulalter vorliegen und die meisten Testverfahren für Kinder unter 5 Jahren kaum geeignet sind.

Prognostische Validität aus klassifikatorischer Sicht

Bei der Sichtung der empirischen Befunde zu den Möglichkeiten der Frühprognose von intellektuellen Fähigkeiten haben wir argumentiert, dass auf der Basis mittlerer Korrelationen zwischen Prädiktorvariablen und späterer Intelligenz keine zuverlässige Vorhersage einer Hochbegabung möglich ist. Entscheidend ist nämlich, mit welcher Wahrscheinlichkeit das prognostizierte Auftreten einer Hochbegabung eintritt oder nicht. Dies lässt sich mit einem klassifikatorischen Ansatz abschätzen, bei dem die Wahrscheinlichkeit für eine zukünftige Ausprägung eines Merkmals wie Hochbegabung durch ein diagnostisches Verfahren eingeschätzt wird. Dazu wird eine inhaltlich begründete Merkmalsgrenze (cut-off-Wert) gewählt, die festlegt, ab welcher Ausprägung des Prädiktors mit dem zu prognostizierendem Merkmal zu rechnen ist. Bei Vorliegen entsprechender Daten lassen sich so alle Personen einem von vier Fällen zuordnen: die Prognose einer Hochbegabung trifft zu (wahr Positive), es wurde fälschlicherweise eine Hochbegabung prognostiziert (falsch Positive), die Prognose einer Nicht-Hochbegabung trifft zu (wahr Negative), es wurde fälschlicherweise eine Nicht-Hochbegabung prognostiziert (falsch Negative).

Die prognostische Validität eines Prädiktors ist dann gegeben, wenn keine falsche Vorhersagen auftreten, also falsch Positive (Fehler erster Art, α-Fehler) oder falsch Negative (Fehler zweiter Art, β-Fehler) ausbleiben. Das ist allerdings unrealistisch, so dass es in der Regel darum geht, beide Fehlerarten so klein wie möglich zu halten. Die beiden Fehlerarten sind nicht unabhängig voneinander, denn meist geht die Reduktion der einen Fehlerart mit der Erhöhung der anderen einher.

Nun kommt erschwerend hinzu, dass die Basisquote im Falle der Hochbegabung extrem klein ist, da nur mit etwa 2 Prozent Hochbegabten in einer repräsentativen Stichprobe zu rechnen ist. In Anlehnung an einen Vorschlag von Marx (1992) lässt sich die Vorhersagegüte eines Prädiktors empirisch über den relativen Anstieg der Trefferquote gegenüber der Zufallstrefferquote (RATZ-Index) bestimmen. Der Index setzt den Anstieg der Gesamttrefferquote gegenüber der Zufallstrefferquote (Differenz zwischen der Gesamt- und Zufallstrefferquote) ins Verhältnis zum maximal möglichen Anstieg (Differenz der Maximal- und Zufallstrefferquote). Marx (1992) hat vorgeschlagen, dass RATZ-Werte ab 66% und größer als sehr gut gelten, Werte zwischen 34% und 66% als gut, und Werte kleiner als 34% als ungenügend für die Individualprognose zu bewerten sind.

Man hat aufgrund der geringen Basisquote schlechte Chancen, wirklich Hochbegabte zu identifizieren (vgl. Tabelle 1). Der größere Anteil der Begabten bleibt in der Regel unentdeckt. Selbst mit den besten der bisher empirisch eingesetzten Prädiktorvariablen ließen sich nicht mehr als etwa 35% der Hochbegabten frühzeitig identifizieren, und die Auswahl wäre dann kaum besser als eine Zufallsauswahl. Unter der Annahme einer utopisch hohen prognostischen Validität von .83 würde man immerhin etwa 90% identifizieren können. Die aktuell verfügbaren Instrumente sind jedoch weit von diesem Wert entfernt und durch die geringe Basisquote hat man immer mit einer relativ hohen Anzahl an nicht Hochbegabten in der Auswahl zu rechnen.

Tabelle 1
Varianten klassifikatorischer Vorhersage-Güte der Hochbegabung auf der Basis der günstigsten empirisch gezeigten regressionsanalytischen Prädiktionswerte nach McCall & Carriger (1993)

a) Annahmen:
Selektionsquote 2% (strenges Kriterium, entspricht der Basisquote)
Prognostische Validität = .36 (vgl. McCall & Carriger, 1993)
Basisquote 2% (entspricht den geläufigen Hochbegabungsdefinitionen)

	Ausgewählte (2)	Nicht Ausgewählte (98)	
Hochbegabte (2)	0.2	1.8	Nur 10% ausgeschöpft
Nicht Hochbegabte (98)	1.8	96.2	
	90% Falsch Positiv		RATZ = 0.08

b) Annahmen:
Selektionsquote 10% (liberales Kriterium)
Prognostische Validität = .36 (vgl. McCall & Carriger, 1993)
Basisquote 2% (entspricht den geläufigen Hochbegabungsdefinitionen)

	Ausgewählte (10)	Nicht Ausgewählte (90)	
Hochbegabte (2)	0.7	1.3	Nur 35% ausgeschöpft
Nicht Hochbegabte (98)	9.3	88.7	
	93% Falsch Positiv		RATZ = 0.28

c) Annahmen:
Selektionsquote 10% (liberales Kriterium)
Prognostische Validität = .83 (entspricht der Retest-Reliabilität bei Erwachsenen)
Basisquote 2% (entspricht den geläufigen Hochbegabungsdefinitionen)

	Ausgewählte (10)	Nicht Ausgewählte (90)	
Hochbegabte (2)	1.8	0.2	90% ausgeschöpft
Nicht Hochbegabte (98)	8.2	88.8	
	82% Falsch Positiv		RATZ = 0.89

Das dynamische Modell der Intelligenz

Die empirischen Befunde zu den Möglichkeiten der Frühprognose von intellektuellen Fähigkeiten ergeben insgesamt ein eher enttäuschendes Bild. Zwar zeigen verschiedene Längsschnittstudien, dass eine gewisse kognitive Kontinuität bzw. eine überzufällige differentielle Entwicklungsstabilität für viele basale kognitive Merkmale vorhanden ist (Alloway & Alloway, 2010; Fagan et al., 2007; Rose et al., 2012), jedoch sind die Korrelationen im Längsschnitt zu gering für eine zuverlässige Prognose von Talent oder für eine frühe Identifikation von Hochbegabten (Perleth et al., 2000; Shapiro et al., 1989). Drei Befunde sind besonders markant (McCall & Carriger, 1993; Sattler, 1988): Erstens ist die Retest-Reliabilität von Tests für Säuglinge und Kleinkinder eher mangelhaft und wird schwächer je höher das Intervall zwischen den Messungen liegt; zweitens wird die prognostische Validität der Verfahren immer schwächer je weiter man in die Kindheit zurück geht; drittens korrelieren verschiedene Verfahren zur Erfassung kognitiver Leistungsniveaus in jungen Jahren eher gering miteinander. Doch warum ist das so?

Vor nicht allzu langer Zeit entwickelten van der Maas et al. (2006) ein dynamisches Modell der allgemeinen Intelligenz (g), das diese Befunde plausibel erklären kann. Das dynamische Modell basiert auf der Annahme, dass kognitive Prozesse (wie zum Beispiel die verbale Verarbeitung oder das Kurzzeitgedächtnis) unabhängig voneinander sind und sich folglich auch zunächst unabhängig voneinander entwickeln. Man beachte, dass diese Annahme fundamental von dem klassischen g-Faktor-Modell der Intelligenz abweicht. Dies beinhaltet nämlich in der Regel die Idee einer einzelnen kognitiven Entität als Grundlage für das gesamte kognitive System: zum Beispiel Speed (Jensen, 1993) oder Arbeitsgedächtnis (Kyllonen & Christal, 1990).

Das Besondere am dynamischen Modell von van der Maas et al. (2006) ist die Annahme, dass einzelne kognitive Prozesse im Entwicklungsverlauf zunehmend miteinander in Interaktion treten, weil bei bestimmten Anforderungen mehrere Prozesse in Anspruch genommen werden. Man spricht von Mutualismus, wenn sich mehrere kognitive Prozesse ergänzen. Man kann sich zum Beispiel vorstellen, dass bei der Dekodierung von Kommunikation sowohl verbale (auditive Verarbeitung) als auch nonverbale Informationen (visuelle Verarbeitung) integriert werden. Dieser Mutualismus führt dann dazu, dass die Leistungsniveaus der einzelnen kognitiven Prozesse und Bereiche zunehmend miteinander korrelieren.

Es lässt sich mathematisch-formal und in Computersimulationen zeigen, dass die Leistungsniveaus am Ende der Entwicklung unabhängig von den Startwerten und von der Entwicklungsgeschwindigkeit der einzelnen Bereiche sind. Das bedeutet, dass die kognitive Leistungsfähigkeit im Erwachsenenalter weitgehend unabhängig von individuellen Differenzen im Säuglingsalter ist. Viel mehr ist das Endniveau abhängig vom Mutualismus, also von der Stärke der gegenseitigen Interaktion der kognitiven Subprozesse und Bereiche. Dabei müssen nicht mal alle Prozesse miteinander in positiver Interaktion stehen und auch die Stärke der Kooperation zwischen Prozessen kann variieren. Wichtig ist nur, dass der interaktive Austausch im Mittel positiv ist, dass also kognitive Prozesse im Mittel eher davon profitieren (sich also schneller entwickeln), wenn sie mit anderen Prozessen kooperieren.

Die Annahme der Unabhängigkeit von Startwerten und Entwicklungsgeschwindigkeit bringt weiterhin mit sich, dass einzelne Prozesse zu Beginn der Entwicklung noch wenig oder gar nicht miteinander korrelieren. Daraus resultiert, dass sich der psychometrische g-Faktor erst im späteren Entwicklungsverlauf herauskristallisiert, also eine Integration der kognitiven Prozesse stattfindet. Diese Annahme ist gut vereinbar mit empirischen Befunden von Rose, Feldman, and Jankowski (2005). Dort zeigt sich, dass die Struktur der kognitiven Fähigkeiten im Säuglingsalter besser multifaktoriell erklärt werden kann als durch ein g-Faktor-Modell. Das schränkt jedoch die Möglichkeiten ein, aufgrund einzelner kognitiver Merkmale in der frühen Kindheit, komplexe kognitive Merkmale wie die allgemeine Intelligenz oder schulische Leistungsfähigkeit zu prognostizieren. Bessere Vorhersagen lassen sich für einzelne Prozesse wie Gedächtnis, Aufmerksamkeit oder Informationsverarbeitungsgeschwindigkeit machen.

Insgesamt kann das dynamische Modell plausibel erklären, warum einzelne kognitive Tests für Kinder im Säuglingsalter wenig miteinander korrelieren und warum Testergebnisse aus der Kindheit wenig mit Testergebnissen aus der späten Kindheit oder dem Erwachsenenalter korrelieren. Außerdem kann das dynamische Modell erklären, warum kognitive Tests der frühen Kindheit eine geringe Retest-Reliabilität aufweisen. Dazu muss man davon ausgehen, dass kognitive Tests (wie zum Beispiel das Habituationsparadigma) nicht nur einen einzigen kognitiven Prozess erfassen sondern unterschiedliche. Man müsste ferner davon ausgehen, dass aufgrund von Variabilität im Mutualismus unter den Prozessen (van der Maas et al., 2006) an unterschiedlichen Zeitpunkten auch unterschiedliche Prozesse bei der Bewältigung von ein und derselben Aufgabe involviert sind. Diese Annahmen sind nicht bewiesen, jedoch gut vereinbar mit dem dynamischen Modell und könnten die mangelhaften Retest-Reliabilitäten von kognitiven Tests im Säuglings- und Kleinkindalter gut erklären.

The Current Research Program

The previous chapter ended on the notion, that there may not be one single cognitive process responsible for the positive manifold. One important implication of this notion is to focus on the research on encapsulated cognitive processes. One obvious candidate for that is working memory capacity. In recent years, working memory has become of particular interest in the research of human intelligence for two reasons. First, working memory and intelligence are highly correlated. Second, the measurement of working memory has similar psychometric qualities as the measurement of intelligence in terms of reliability and predictive validity.

Measuring Working Memory

One of the first and frequently cited studies in this context was conducted by Kyllonen and Christal (1990) and was an attempt to explain individual differences in intelligence with an information processing model. That is, instead of relying on a mysterious g-factor to explain the positive manifold, they assumed a small network of cognitive processes, at the heart of which lay working memory. Along the way, it was also one of the first attempts to define working memory psychometrically as a latent variable. The tests to measure working memory capacity were selected on theoretical considerations of Baddeley (1986), defining working memory as a limited capacity system responsible for "temporary storage of information that is being processed" (p. 34). This has become the hallmark definition whenever it came to the measurement of working memory and it is found in almost every publication in the field. This is interesting because it is such a narrow definition and entirely focused on cognitive processes, unlike definitions of intelligence that mostly try to capture the versatility of the human mind (e.g., Gottfredson, 1997).

One working memory task in Kyllonen and Christal (1990), for example, was to integrate the information of three consecutive sentences. Each sentence contained information about the relationship of the letters ABCD, such as "A precedes B" or "C precedes D". Sentences could not be revisited, so that the information had to be memorized, and with each consecutive sentence, new information had to be considered and the mental representation of the four letters had to be adjusted accordingly. In another task, subjects were asked to engage in mental arithmetic with variables. Formulas were presented consecutively and could not be revisited, for example "A = B/2" or "B = C – 4". Thus, mental operations had to be performed on the variables and their values updated accordingly. It is relatively clear to see how these tasks require the simultaneous storage and manipulation of old and new information over a short period of time.

In a series of four studies, Kyllonen and Christal (1990) demonstrated correlations between the latent variables for working memory and reasoning (a close relative to general fluid intelligence) that were consistently beyond .80. This was somewhat surprising because the tasks used to measure reasoning seemed more diverse in their mental requirements. In one reasoning task, for example, participants were presented four sets, each containing three digits (e.g., “282, 848, 244, 566”). The digits in each set were governed by a hidden rule so they had something in common (e.g., all even digits). But one of the number sets violated that rule (e.g., contained an odd digit), and participants were asked to identify the one set that did not follow the rule. At least superficially, it is hardly recognizable how such a task would rely on memory because all information is readily available and nothing needs to be remembered or recalled. Another reasoning task consisted of verbal analogies, like for example “DISTANCE:MILE::VOLUME:?”, whereas the correct analog should be selected from a given list (“LITRE, BOTTLE, WATER”). Again, it is hard to see how this would rely on working memory and rather seems to rely on verbal knowledge. Based on the observation of very high correlations, Kyllonen and Christal argued that reasoning ability is mostly determined by working memory capacity and conjectured that “working-memory capacity affects success across the various component stages of reasoning tasks” (p. 427). They basically claimed to have found the source process accountable for the g-factor.

However, some other tasks that Kyllonen and Christal (1990) used to measure reasoning can, in retrospect, be assumed to be measures of working memory capacity. There was, for example, one reasoning test that contained arithmetic word problems of the form: “Pat put in a total of 16½ hours on a job during 5 days of the past week. How long is Pat's average workday?” (Department of Defense, 1984, as cited in Kyllonen & Christal, 1990, p. 394) This kind of mental arithmetic would fit the definition of temporary storage and processing, in that it requires performing the mathematical operations and storing intermediary results before arriving at the final result. Other tasks, attributed to reasoning, were linear syllogism problems such as: “Dick is better than Pete, John is worse than Pete, Who is best?” Here too, the clauses can be assumed to be processed consecutively, so that the relation between John and Pete is processed and integrated into the relation between Dick and Pete. This theoretical overlap of latent variable indicators could very well explain why the correlations obtained in confirmatory factor analysis models were of such magnitude. Indeed, this study inspired a host of subsequent studies regarding the relationship between intelligence and working memory and, although all of them found a considerable correlation, the correlations were never as high as suggested by Kyllonen and Christal.

Fifteen years later, so much research had accumulated on the topic that Ackerman et al. (2005) asked “working memory and intelligence: the same or different constructs” (p. 30) and tried to find an answer in a comprehensive review and meta-analysis. They estimated the meta-analytical average correlation at around .50 and concluded that, although both constructs share considerable amount of variance, the two are dissociable. In a comment on the very same paper, Oberauer, Schulze, Wilhelm, and Süß (2005) pointed out that it is not only important to distinguish working memory and intelligence by means of correlation, but that it is even more important to highlight conceptual differences:

> From a theoretical point of view, there is no reason to assume that [working memory capacity] is the same as *g*. By definition, *g* is conceptually opaque – it is the common variance of a set of tasks that happened to be constructed and used by intelligence researchers over a century. It reflects no explicit theoretical concept, and hence there is no theory-based procedure for measuring it. Rather, g reflects a mixture of the mostly implicit theories of intelligence various researchers have endorsed and their intuitions about ways to test it. (p. 64)

This quotation taps into the key difference between working memory and intelligence. As pointed out in the introduction, the scientific pursuit of the measurement of human intelligence started with the tests. It was only after there was already a considerable diversity in measurement approaches when Spearman (1904) identified the g-factor. There was no theory about the mechanism, and theories to explain the g-factor were proposed and investigated ex post facto. The working memory construct, on the other hand, resulted from theories and research in the field of memory and inspired measurement methods that were strictly confined to the theoretical cognitive processes.

One influential working memory model was partly derived from the observation that multitasking, although with some detriment to performance, is possible. Hitch and Baddeley (1976) asked participants to perform two tasks simultaneously. The main task was, what Hitch and Baddeley called a verbal reasoning task. It consisted of a verbal statement describing the relationship between two letters (e.g., "A follows B") and an arrangement of two letters that could either be consistent with the statement ("BA") or inconsistent ("AB"). The subjects were asked to respond as quickly as possible. Note that this task is arguably less complex than the inductive reasoning tasks that are the subject of this dissertation, and Hitch and Baddeley (1976) report that error rates were usually well below 10%. Nonetheless, this task should draw on resources in working memory by requiring the translation of the verbal information into a mental representation that can finally be compared against the visual information to make a congruence judgement. This task was combined with various versions of a secondary task that should compete for short-term storage capacity. A rather simple version of the secondary task required to verbally repeat the simple word "the" out loud at a high rate. A more complex version required to repeat a sequence of six random digits out loud. The results of these experiments revealed that most of the simple load tasks had almost no impact on verbal reasoning, and even the rather complex load tasks impaired performance and latencies far less than expected. Hence, Hitch and Baddeley assumed that short-term memory not only has a limited storage capacity, but also a processing mechanism that would account for performance on concurrent tasks. The multicomponent model therefore assumes two temporary storage systems, respectively for verbal and visual information, which are both controlled by the central executive, responsible for focusing, dividing, and switching attention. As a whole, this system is responsible for ongoing information processing and can account for findings in reasoning, learning, and comprehension (see Baddeley, 2003).

Although, there are still many unknowns in this field of research, especially concerning the central executive system (Vandierendonck, 2016), the precise theoretical framework lead to the development of a small set of tasks that were repeatedly used to measure working memory capacity. Prominent among those are complex span tasks, grounded on the notion that simultaneous storage and processing is a key aspect of information processing in working memory (Daneman & Carpenter, 1980). A systematic comparison and overview of some of these tasks (reading span, operation span, and counting span) was provided by Conway et al. (2005). All complex span tasks follow a similar procedure and involve two subtasks. For example, the reading span requires test-takers to read a sentence and judge its validity (true or false). This makes for the processing component. The storage component lies in the requirement to memorize the last word of the sentence for later recall. The operation span task works essentially the same way, except that math equations are to be verified as the processing component. The processing component in the counting span task requires test-takers to count stimuli with a certain feature (e.g., counting all red circles). The storage component lies in the requirement to memorize said counts for later recall. The relevant dependent variable is usually the storage component, whereas the processing component is considered to be very easy with generally few mistakes. The processing component is assumed to interfere with active rehearsal of the storage component. Hence, the task should provide an estimate of the amount of information that can be actively maintained during distraction.

There are other tasks to measure working memory capacity and Cowan et al. (2005) pointed out some of the difficulties in interpreting dual task scores. After reviewing and comparing some prominent working memory measures, they concluded that dual tasks are not really necessary to obtain a good estimation of a person's working memory capacity for individual differences research. They propose to measure what they call the "scope of attention", which is the amount of information that a person can attend to at a given point in time.

One task to measure the scope of attention is the running memory span, where a constant stream of digits (or other stimuli) is presented to the test-taker. The stream will disrupt at some random point and test-takers are then asked to recall all the digits up the point of disruption in forward order. The assumption is that the items of the list are retrieved from a stream of sensory memory that forms automatically from active attention and perception during number presentation. Another example is the visual array task which presents consecutively two arrays of colored squares that should be compared. One of the squares is marked on the second array and participants are asked to indicate whether it had the same or different color in the previous array. Again, the assumption is that the array of squares will be retrieved from a visual sensory memory trace that forms from perception, and that there cannot be enough time in the brief interval between arrays to rehearse verbally or form chunks.

The idea with these kinds of tasks is that sensory memory has a comparably large capacity but is very short-lived. Thus, information from this short-lived memory needs to be transferred into the capacity-limited scope of attention so it can be recalled (Cowan, Fristoe, Elliott, Brunner, & Saults, 2006). Although this theory is different from classical storage and processing theories of working memory, the general idea is the same: Limited capacity affects online information processing.

Working Memory vs. Matrix Reasoning

In the previous section, I have established that working memory and intelligence are closely related, however there are differences. It is not easy to work those out, mostly because intelligence is such a diverse construct. Take the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 2009) as an example. The test battery has a total of 14 subtests which can be very different. One of those is the “digit span” test, which is essentially a short-term memory test for digit recall and part of the working memory subscale in the WAIS. So on this specific task, there is literally no difference between working memory and intelligence. Then there is the verbal comprehension subscale, which consists of verbal knowledge subtests like “vocabulary” and “comprehension”. It is fair to assume that those cover long-term memory contents (i.e., crystallized intelligence) but should not be primarily dependent on working memory capacity.

Another subtest requires test-takers to complete a matrix of figural stimuli that are arranged according to some hidden rule. This kind of task is of particular relevance in this dissertation because of the requirement to detect hidden rules, which is the main characteristic of inductive reasoning. Inductive reasoning means to recognize regularity and to infer a general rule from a particular example and to generalize this rule to other conditions. In the specific case of matrix reasoning, this means to recognize how figural elements in the visible part of the matrix are related and to infer how the missing piece should look like. Matrix reasoning tests have somewhat of a special standing in intelligence research, and this is especially true for Raven’s matrices (Raven, Raven, & Court, 1998). Spearman (1938) already remarked on an early version of the test that he regarded it very suitable to measure intelligence. Others have demonstrated that Raven’s matrices share a large amount of variance with the g-factor (Marshalek, Lohman, & Snow, 1983) and Carroll (1993) noted: “Our evidence suggests that the Progressive Matrix test is a good measure of *g* and of the second-stratum factor 2F, but the degree to which this test measures first-order factors I and VZ is not clear“ (p. 696). The test is also very easy to administer and to score, which is probably why it is very popular among researchers and very frequently used in all sorts of studies to provide an estimate of intelligence.

Therefore, the study of the relationship between working memory and intelligence is oftentimes a study of the relationship between working memory and matrix reasoning. Even in studies that work with a g-factor definition of intelligence, matrix reasoning tasks are almost certainly part of the test battery. One of such studies was reported by Conway, Cowan, Bunting, Therriault, and Minkoff (2002) where, in addition to Raven's matrices, Cattell's Culture Fair Test was used as a marker of fluid intelligence. The Cattell test comprises of four subtests, one of which is a matrix reasoning test as well, and the others are mostly variants of inductive reasoning tests as well. Conway et al. reported results from structural equation models that estimated the relevance of processing speed, working memory, and short-term memory as predictors. Notably, working memory tasks were allowed to load on the short-term memory factor to draw storage variance and leave the processing part. In this model, short-term memory and processing speed were no significant predictors but working memory was a strong predictor.

Conway et al. argued that controlled attention, responsible for information processing in the face of distraction, would be most relevant because inductive reasoning would bring about an interplay of discovery and maintenance of rules: “In order to solve difficult matrix problems, one must discover a rule and then maintain that rule while searching to discover a second rule and so on” (p. 179). Since this account would involve the temporary storage of discovered rules and intermediate results, it is surprising that Conway et al. could not estimate a significant relationship with the storage factor (but see Martínez et al., 2011 for a different finding).

Conway et al. (2002) also discussed the possible role of strategy use for the relationship between working memory and intelligence. They argue that previous studies have found that individual differences in strategy use can account for performance on both, working memory tasks and intelligence tests. Hence, the common ability to both tasks would “involve the recognition and successful execution of particular strategies” (p. 179).

Unsworth and Engle (2005) hypothesized that working memory capacity would be important in matrix reasoning because test-takers need to store a certain amount of relationships on each problem. Thus, they investigated the correlation between single item response accuracies and working memory capacity. However, they found that correlations were constant across item difficulty levels, and could not be accounted for by the amount of rules underlying each problem. Unsworth and Engle discuss that short-term storage must be of little importance but “it is the central executive component of the working memory system that is important on both working memory span tasks and tasks of fluid abilities” (p. 78).

Wiley, Jarosz, Cushen, and Colflesh (2011) tested the hypothesis that the discovery of new rules in matrix reasoning tests can be mostly accounted for by working memory capacity. Their analyses revealed that all items that make use of a new rule or new combination of rules in the Raven test, were stronger correlated with working memory capacity than the other items. They argue that “previously learned and used rules may interfere with performance when a new rule combination is needed” (p. 261). Hence, the ability to control the focus of attention (executive functioning) would be relevant in ignoring irrelevant solution approaches.

Harrison, Shipstead, and Engle (2014) failed to replicate the findings of Wiley et al. (2011). They constructed a special composition of matrix reasoning items and presented two parallel items, with the same underlying rules, in consecutive order. Hence, they could compare how very similar items correlate with working memory, depending on whether they were presented first or second. The pattern of results was contradictory to Wiley et al.’s findings in that repeated-rule problems correlated significantly higher with working memory than first-rule problems. The authors argue that “one of the reasons that [working memory capacity] is correlated with Raven’s problems is possibly that subjects with high [working memory capacity] are able to retrieve solutions from previous Raven’s problems to solve the current problem” (p. 394). This explanation puts the finger back on the storage component of the working memory system as a major source of variance in matrix reasoning; however Harrison et al. (2014) acknowledge that there could be multiple mechanisms that account for this shared variance.

It is interesting to note that none of the accounts for the relationship between working memory capacity and matrix reasoning make explicit assumptions about the source of rule discovery. Most of them acknowledge that rules play some role, either because they need to be retained in storage or need suppressed as distractions. This raises the question as to the nature of the rule induction process itself, which is the key characteristic of matrix reasoning tests as a subcategory of inductive reasoning tests and general fluid intelligence tests. Thus, the following experiments are an attempt to explore this aspect of matrix reasoning in more detail. The hypothesis is that working memory can in itself not account for individual differences in the ability to discover rules. This is thought to be a key difference between working memory and inductive reasoning and may as such represent a conceptual difference between working memory and intelligence. The research reviewed herein suggests that working memory capacity is critical for almost all measures of fluid intelligence, however inductive reasoning should also be affected by some additional process that can account for the ability to come up unknown rules and relations in matrix reasoning.

Part 2: How Knowing the Rules Affects Solving the Raven Advanced Progressive Matrices Test

Publication Note

The following part was first submitted on 15.1.2013, submitted after revision on 2.9.2014, and published as: Loesche, P., Wiley, J., & Hasselhorn, M. (2015). How knowing the rules affects solving the Raven Advanced Progressive Matrices Test. *Intelligence,* 48, 58-75. doi: 10.1016/j.intell.2014.10.004

Abstract

The solution process underlying Raven's Advanced Progressive Matrices (RAPM) has been conceptualized to consist of two subprocesses: rule induction and goal management. Past research has also found a strong relation between measures of working memory capacity and performance on RAPM. The present research attempted to test whether the goal management subprocess is responsible for the relation between working memory capacity to RAPM, using a paradigm where the rules necessary to solve the problems were given to subjects, assuming it would render rule induction unnecessary. Three experiments revealed that working memory capacity was still strongly related to RAPM performance in the given-rules condition, while in two experiments the correlation in the given-rules condition was significantly higher than in the no-rules condition. Experiment 4 revealed that giving the rules affected problem solving behavior. Evidence from eye tracking protocols suggested that participants in the given-rules condition were more likely to approach the problems with a constructive matching strategy. Two possible mechanisms are discussed that could both explain why providing participants with the rules might increase the relation between working memory capacity and RAPM performance.

Introduction

There is reasonable evidence that working memory capacity plays a crucial role in human intelligence. Most studies that have contributed to this finding follow a methodology where each construct is operationalized with some representative tests and then the correlational pattern is subject to analysis, utilizing latent factor modeling. In most cases some other constructs are also being taken into account, like short term memory, processing speed, or long term memory. The latent variable correlations or factor weights describing the relation between working memory capacity and intelligence are usually quite substantial even when other variables are taken into account (see for example, Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004; Conway et al., 2002; Engle, Tuholski, Laughlin, & Conway, 1999; Kyllonen & Christal, 1990; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002). However, this correlational approach has its limits: From the studies mentioned above we can conclude that there is some substantial association and that working memory capacity plays a bigger role than other cognitive resources, but we cannot tell exactly where the relation stems from.

Ackerman et al. (2005) stated that "resolution of the question of how and how much working memory and intelligence are related ultimately requires additional research" (p. 52). Oberauer et al. (2005) have argued that the "distinction between these constructs does not hinge on the size of the correlation but on a qualitative difference…" (p.63). This leads us to suggest that with a plain correlational approach we cannot conclude exactly how basic cognitive resources like working memory capacity are involved in the processing of intelligence tasks and to what degree. For this reason, the present article shall explore the borders of the relation between intelligence and working memory capacity by combining the prevailing correlational approach with an experimental methodology. The central question is: Where lies the common link between working memory capacity and intelligence?

The Two-Process-Theory of Inductive Matrix Reasoning

We will approach this question by examining the solution process of the Raven Advanced Progressive Matrices (RAPM) test (Raven et al., 1998). Typical RAPM items require test-takers to analyze figural elements in a matrix in order to select the correct solution out of eight response alternatives (see Figure 1). According to Carroll (1993), tasks of this kind form a good representation of the general fluid intelligence factor (gf) which he describes as being "concerned with basic processes of reasoning and other mental activities that depend only minimally on learning and acculturation" (p. 624). His analysis suggests that tasks of this kind load reasonably high on the g-factor, the gf-factor, and various reasoning factors and indeed, matrix reasoning tasks are included in numerous prominent test batteries for intellectual assessment, including the WAIS-IV and the SB5. The solution process in inductive reasoning tasks of this sort has been subject to analysis in previous studies and there is some understanding of the processes involved.

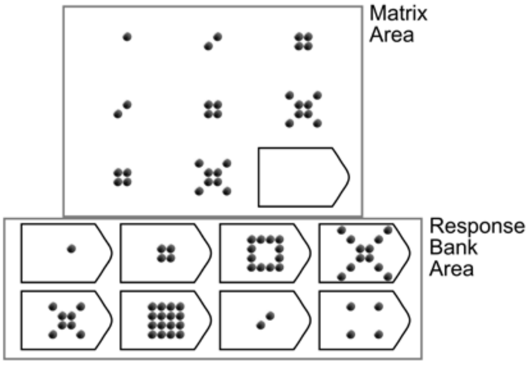


Figure 1. Prototypical problem of Raven's APM with two areas of interest designated.

Carpenter, Just, and Shell (1990) have contributed substantially in this vein in their approach to simulate successful human performance on RAPM with a computer program. In order to reach this goal, they performed a task analysis of problem solving behavior in RAPM using techniques like eye tracking and think aloud protocols. Using this approach, they articulated two subprocesses that distinguish among higher and lower scoring individuals: the ability to induce abstract relations and the ability to dynamically manage a large set of problem solving goals in working memory.

Rule induction refers to the process of finding abstract relations among the elements in the figural matrices and concluding which rules guide these relations. Based on their research, Carpenter et al. (1990) postulate a taxonomy of five different types of rules that would be sufficient to describe the relation among elements for most of the items in RAPM. They describe the process of finding these correspondences like a trial-and-error method, where a subject tries to identify some elements in the matrix with a rule, and if it leads to a dead end, tries a different rule or different elements. According to their analyses, correspondence finding involves the decomposition of the figures into their composing elements and comparing them pairwise; furthermore the process is proposed to be sequential, which means that one rule is induced at a time.

Goal management refers to the process of setting and monitoring goals and subgoals during problem solving. The main goal is evidently to solve the problem, but in order to reach this, subgoals need to be created, like finding a connection among certain elements (i.e. correspondence finding). This process involves the association of the figural elements in the matrix with certain subgoals. Also, the process involves monitoring the relations found and keeping them present in working memory. That is, once a relation is regarded as valid it has to be maintained before the search for further rules among other elements can continue.

Carpenter et al. (1990) offered two results that suggest that goal management processes are largely responsible for successful performance on RAPM. First, in their Study 1A, they reported a correlation of -.57 between the number of rule tokens required to solve each problem and its solution rate. On the basis of these results they argued that “the presence of a larger number of rule tokens taxes not so much the processes that induce the rules, but the goal-management processes that are required to construct, execute, and maintain a mental plan of action during the solution of those problems containing multiple rule tokens as well as difficult correspondence finding” (p. 410). Hence they argued that as the number of rules increases, the demand placed on working memory capacity increases as well. Second, in Study 2, they taught participants how to solve another problem solving task, the Tower of Hanoi, using a recursion strategy, and showed that performance on that specific version of the task, where the need to induce the recursion strategy was removed, was also highly correlated with performance RAPM (*r* =.77). Given the high relation between this modified Tower of Hanoi task and performance on RAPM, which they attribute to the need for goal management on both tasks, they raise the question whether there is any need to postulate other processes, such as abstraction or inductive ability, as additional sources of individual differences in the Raven test.

The Role of Working Memory Capacity in RAPM

The working memory concept originated from the notion that complex cognitive tasks need information readily accessible, and it was further put forward with the distinction between primary and secondary memory (Berti, 2010). The observation that it is actually possible to combine two relatively complex tasks without any disastrous detriment in performance on either task let to the conclusion that there had to be some sort of managing system (the central executive) that is responsible for the coordination of simultaneous processes, especially when the capacity limit for short term storage is reached (Baddeley & Hitch, 2007). As an individual differences measure, working memory capacity can be seen either as a measurement of the amount of information that a person can store and retrieve in the face of a competing task, or alternatively, as the ability to make the most effective use of this system via the use of attentional control or executive functioning (Conway et al., 2005; Cowan et al., 2005; Kane, Conway, Hambrick, & Engle, 2007).

Several previous studies of RAPM performance have suggested that item characteristics, like the number of elements and rules, affect item difficulty by placing demands on working memory (Embretson, 1998; Primi, 2001). Arguably, the sheer amount of elements and rules that need to be handled while solving an item would exceed the storage capacity of working memory. Still, working memory capacity has not been assessed directly in these studies and, to the contrary, several studies that have assessed working memory capacity have failed to find a relation with item difficulty (Salthouse & Pink, 2008; Unsworth & Engle, 2005; Wiley et al., 2011). For example, Unsworth and Engle (2005) showed that item difficulty is not at all related to working memory capacity. They found that performance on individual items is rather constantly correlated with working memory capacity. Furthermore, Salthouse and Pink (2008) found out that the correlation between memory span and gf is fairly independent from the list length in the memory tasks. Similarly, several researchers have also been unable to find relations between the number of rules or rule tokens and working memory capacity. If more cognitive load is put on working memory and goal maintenance processes due to increased numbers of rules, then the relation between working memory capacity and RAPM performance should increase as the number of rules required to solve the problems increase. However, several studies have reported that the relation between working memory capacity and RAPM remains constant across items regardless of the number of rules or rule tokens that they require (Unsworth & Engle, 2005; Wiley et al., 2011). Hence, although early work provided support for the hypothesis that the relation between working memory capacity and RAPM is largely due to goal management processes, more recent research suggests the role of goal management in explaining the relation between working memory capacity to RAPM performance is still unclear.

The Current Paradigm

As noted earlier, Carpenter et al. (1990) suggested that goal management is the crucial process in the RAPM solution process, however, the rule induction process in their simulations works presumably differently from an actual human cognitive process. Their computer program was designed in a way that it searched for applicable rules to solve the problem at hand from a finite set of rules, formed by the five rules from their taxonomy. The program does not account for the possibility that someone might take a completely different approach which may lead to a dead end, or, by mere chance, also to a correct solution. For a human being, the rules stem from a potentially greater population of solution strategies. A human being who has never encountered the problems before has to come up with an idea about how to approach the problem in the first place. Verguts, De Boeck, and Maris (1999) describe this step like sampling rules from an urn until all element relations in the problem can be accounted for.

We asked ourselves: What would happen if the set size of the urn would be reduced to the number of rules that are actually applicable to the problems? What if humans already know the rules, as was the case for Carpenter et al.'s computer programs? More specifically, we were interested to what degree working memory capacity is involved in the sampling of new rules. The research that has linked working memory to gf has, to this point, mainly focused on the part that is not involved in generating rules. The prevailing accounts for the correlation envision some sort of information processing that involves storage, maintenance, inhibition, supervision, attention, or updating, but none of these accounts can explain how a mental representation of a rule or abstract relationship is actually formed. Furthermore, it lies in the nature of working memory tasks that they are free of inductive processes. That is, in typical working memory tasks participants are fully informed about the task and about the relation of the task material to a correct response, so that performance is solely limited by capacity. In our view, this aspect is fundamentally different from intelligence tests like RAPM, where the connection among stimuli is unknown to the subject.

Over a course of four experiments we wanted to shed some new light on the relationship between working memory capacity and rule induction by introducing a new paradigm that involves teaching the rules that would be necessary to solve problems from Raven's APM. We predicted to find an increased correlation between RAPM and working memory capacity when the rules are known in Experiment 1. We further predicted to find the opposite pattern for the correlation between RAPM and measures of rule induction and productive thinking in Experiments 2 and 3. Finally, in Experiment 4 we predicted to find different patterns in eye movement behavior while solving RAPM problems, depending on whether the rules are known or not.

Experiment 1

As mentioned before, we know that there is some decent correlation between RAPM and measures of working memory capacity. However, it is unclear how individual differences in working memory capacity are affecting performance on RAPM. To test whether working memory capacity contributes to performance on RAPM via its influence on goal maintenance alone, we conceived of an experimental manipulation which eliminates the need to induce rules during the solution process of RAPM: Teaching the rules necessary to solve the problems even before test-takers tackle the problems.

This manipulation involves teaching participants five rules, first developed by Carpenter et al. (1990), and having them solve a subset of the items that can be solved using those rules (see Table 2 for a description). The assumption is that if the test-takers know the rule taxonomy, they would simply have to recall the rules and check if any of the rules are applicable. They would not have to rely on rule generation or hypothesis formation, meaning that any relation between working memory capacity and RAPM performance in this case should be due to goal maintenance processes.

On the contrary, if goal maintenance processes do not play a unique role in the relation of working memory capacity and RAPM performance (for example if rule induction is largely responsible for the relation), then the relation between working memory capacity and RAPM performance should be decreased when only goal maintenance is required. Thus, the amount of variance in RAPM predicted by working memory capacity would be lower than in the control group where rule induction is still required.

Table 2
Taxonomy of Rules (Based on Carpenter et al., 1990)

Original rule name	**Description**	**Rephrased rule name**
Constant in a row	The same value occurs throughout a row, but changes down a column.	Always the same
Quantitative pairwise progression	A quantitative increment or decrement occurs between adjacent entries in an attribute such as size, position, or number.	Progress
Figure addition or subtraction	A figure from one column is added to (juxtaposed or superimposed) or subtracted from another figure to produce the third.	Plus Minus
Distribution of three values	Three values from a categorical attribute (such as figure type) are distributed through a row.	One of each
Distribution of two values	Two values from a categorical attribute are distributed through a row; the third value is null.	--

Method

Redraft of the Rule Taxonomy

Since the rule taxonomy developed by Carpenter et al. (1990) consists of somewhat long and technical terms like “quantitative pairwise progression” and since we decided to work with children as participants (for reasons explained later), we used only a subset of the problems that could be solved with five of the rules, and rephrased some of the rule names (see Table 2). This was intended to make it easier for participants to remember and understand the rules. The “constant in a row” rule was rephrased to *always the same*, the “quantitative pairwise progression” rule was rephrased to *progress*, the “figure addition or subtraction” rule was split into two corresponding rules named *plus* and *minus*, and the “distribution of three values” rule was rephrased to *one of each*. We dropped the “distribution of two values” rule because most items where this rule is applicable are also solvable via *one of each* or *plus* or *minus*. Additionally, this omission kept the instructions shorter, which was intended to make it easier to remember all rules.

Working Memory Assessment

All task materials for working memory assessment were adapted from Vock and Holling (2008), where they proved to be appropriate for use with children from 8 to 13 years of age. The tasks were chosen to represent each of three possible task modalities (verbal, spatial, and numerical).

The first task was a spatial working memory task (SWM). In this task a series of 3x3 patterns with white and black squares was presented sequentially, each pattern for 1.5 seconds. Before each series of patterns, an arrow indicated the direction in which these patterns had to be rotated mentally; either 45 degrees to the right or to the left. The length of series increased from 1 to 4. After each series, the participant was required to change the colors of blank 3x3 checker fields to indicate his or her memory of the mentally rotated patterns. There was a 60 second time restriction on the response screen. After that or when the participant pressed an ok-button, the next item was immediately presented. Four practice items preceded the 13 test items. Each correctly recalled pattern was scored with one point divided by the number of patterns on the item (partial credit scoring), for a maximum possible score of 13 points.

The second task was a backward digit span task (BDS). In this task a series of digits between 1 and 9 was presented sequentially, each digit for 1.5 seconds. The length of the series increased from 4 to 7. After each series, the participant was required to enter the series backwards in a textbox. The participants could indicate missing digits with an underscore. There was a 60 second time restriction on the response screen. After that or when the participant pressed an ok-button, the next item was presented immediately. Two practice items preceded the 12 test items. Recalled digits had to be in the correct position within the series. Each correctly recalled digit was scored with one point divided by the number of digits of the item (partial credit scoring), for a maximum possible score of 12 points.

The third task was a verbal span task (VS). In this task, a list of words was presented for 6 seconds. The length of the list increased from 3 to 6 words, each word consisted of no more than two syllables. After each list, distraction tasks followed in which an array of five words was presented on the screen. A category term was placed in the center with four nouns in the corners. Participants were asked to click on the correct word that was a member of the category. The number of distraction tasks alternated between two and three. There was no time restriction on the distraction tasks. Afterwards, a textbox was presented where participants could enter the words they remembered from the list, via keyboard. The participants were instructed that minor spelling and typing errors were not important to scoring. A 90 second time restriction was displayed on the response screen. After that or when the participant pressed an ok-button, the next item was immediately presented. Two practice items preceded the 10 test items. Recalled words had to be in the correct order relative to the other correctly recalled words. Each correctly recalled word was scored one point divided by list length (partial credit scoring). Errors of commission and errors of omission were ignored. The maximum possible score was then 10 points.

In order to have a whole score for working memory performance, a composite working memory task score was calculated by averaging z-scores of spatial working memory and backward digit span for cases that had no missing data on these tasks. For consistency with the other experiments reported here, the verbal span task was omitted in this composite score (which did not affect the pattern of results).

Raven's APM

The fourth task was Raven's Advanced Progressive Matrices (RAPM). The main task was preceded by an instruction video and some practice items. In both groups, the video explained the task. Differences between the two groups were as follows:

The control group received an instruction that very closely followed the manual (Raven et al., 1998). That is, Item 1 from Set 1 was shown and it was explained that one had to infer what kind of piece was missing in the displayed pattern. Two wrong solutions were shown before the right solution was given, and it was explained why they were right or wrong. An explanation was given that one had to look for the underlying rules which might apply from left to right or top to bottom. Then Item 2 was shown and the video gave the participant some time to think for him- or herself; afterwards the correct solution was given. Then again, it was explained that it was important to look out for the principles on which the tasks work, and the participant could practice for 12 minutes on some tasks which would not be scored. Then the 12 items from Set 1 were presented. The length of the instruction video was 3:00 minutes.

The experimental group received an instruction that emphasized the five rules: *Always the same, progress, one of each, plus*, and *minus*. First, participants were informed that the task was to identify the missing piece in the problem from the eight pieces given. Then, four items from Set 1 were shown to exemplify the rules. Item 6 was shown to exemplify the rules *always the same* and *progress*, then Item 7 was shown to exemplify *always the same* and *one of each*, then Item 10 was shown to exemplify *plus*, then Item 12 was shown to exemplify *minus*. Colored animations helped to visualize the important elements for each rule on each item. Furthermore, an explanation stated that the rules would apply from left to right and that the rules would be sufficient to solve any problem in the test. After that, the same four items were presented again and the participants were given 5 minutes time to work on them as practice items which would not be scored. The length of the instruction video was 6:41 minutes.

In the main task, only 26 items from Set 2 were deployed since careful analysis of the Items 15, 18, 19, 20, 25, 30, 31, 33, 35, and 36 indicated that the rules described earlier would not apply in the same way as they would for the other items. For item 15, the rule *plus* does apply, but it applies in different directions for different elements, whereas the instructions in the experimental condition emphasized that participants should look for rules from left to right. For item 18 none of the aforementioned rules apply. A new rule that could be called “morph shape” (Wiley et al., 2011), would apply here, which however would not apply to any other problem in the set. For Items 19, 20, 25, 30, 33, 35 the *plus* and *minus* rules do not apply in their simplest way, instead a specific plus/minus rule would have to be inferred for certain elements. For example for Item 33, one would have to infer something like “same color + same color = increase” and “different color + different color = decrease”. For other items from this list, the plus/minus rule would need some differential consideration of foreground and background, like in Item 20 where blank patterns are always on top. To keep the instructions as simple as possible, these subtleties of the *plus* and *minus* rules were omitted. Note that Items 18 and 19 were also not classified by Carpenter et al. (1990) because “the nature of the rules differed from all others” (p. 408). For Items 31 and 36 the “distribution of two values” rule would have to be employed. However this rule was omitted in the current study for reasons explained earlier.

All 26 included items were presented individually on the computer screen in a similar fashion as the paper-and-pencil version of the test. The participants were required to indicate via button click which of the eight given solutions they thought would be correct. They could freely move through the set of items forward and backward and always change their responses. Whenever they thought they were done with the task, they could just press a button to finish. If they did not finish manually, the task would terminate automatically after a 30 minute time limit (which occurred only with 2 participants).

Procedure

The participants were tested in groups of 5 to 22 individuals (*M* = 10.9, *SD* = 5.3). They had about 1.5 hours time, which was always sufficient to complete all of the tasks. All tasks were presented on computers, which were controlled partly by mouse and partly by touchscreen. The students wore headphones throughout the tasks to receive audio-visual instructions. The setting was either a classroom or a computer lab at the school. After a brief introduction to what the study was about, how many tasks the participants would encounter, and the nature of the tasks, they could start at their own will. After the starting screen where age and gender was inquired, participants got to a screen with buttons for each task. Each task could be started with pressing a button and each task was preceded by a short introduction video. Participants were allowed to take breaks at leisure between tasks. The order of the tasks was fixed, which was accomplished by enabling or disabling the corresponding buttons, based on which tasks had already been completed. The assignment to one of the two experimental groups was accomplished via a built-in random number generator in the computer program.

Pretesting had indicated that participants' working time on the tasks could vary quite strongly. One of the consequences was that students who finished earlier than their classmates may have influenced slower students in their task performance. Thus a dummy task was added as a fifth task, just to keep the quick students busy for a while. This task was a spatial working memory task which did not produce any data.

Sample

Due to the experimental manipulation a ceiling effect was likely to occur on RAPM, thus the target population should be from the lower end of the ability distribution. Accordingly, the decision was to target students from grades 5 to 8, which includes the youngest age group RAPM is applicable to, according to the manual (Raven et al., 1998). The participants were located in 4 different secondary schools in Frankfurt am Main, Germany. Their parents were required to sign an informed consent as a prerequisite for participation. The participants were informed about the voluntary nature of their participation and confidentiality of their responses. Each participating class received a donation of 150€ to the class treasury as a compensatory payment.

The total amount of participants tested for the study was 647. However, there was missing data on all 4 main tasks from 3 participants who were therefore excluded from the analysis. Reasons for missing data were mostly technical issues, like sound, video, mouse, or keyboard crashing. Furthermore, any score equaling zero on any of the main tasks was recoded to missing, assuming there may have been motivation or comprehension issues. The sample for analysis, then, consisted of *N* = 644 secondary school students, from which 316 were randomly assigned to the control group and the remaining 328 to the experimental group. Grade was approximately evenly distributed (grade 5 = 165, grade 6 = 214, grade 7 = 137, grade 8 = 128), as was school level (350 from the highest level, 294 from the second highest level), and gender (319 male, 325 female) among participants. The participants' age ranged from 10 to 16 years (*M* = 12.2, *SD* = 1.3).

Results

First, means and differences in task performance are reported for the two experimental groups, which can be obtained from Table 3. Task performance in the working memory tasks did not differ significantly between groups (*t*s < 1.08, *p*s > .28) and the variances of the working memory tasks did not differ significantly between groups (Levene's *F*s < 1.53, *p*s > .28). Also, the correlations among the three working memory tasks did not differ significantly between experimental groups (see Table 4). This suggests that random assignment resulted in an evenly distributed working memory capacity profile in both experimental groups.

Second, task performance on RAPM was significantly better in the given-rules group than in the control group, $t(624) = 6.40$, $p < .01$, $d = .52$. On average, participants in the experimental group were able to solve about 2.3 items more, due to knowing the rules underlying the problems. There was no ceiling effect, as indicated by skew = -.03 and maximum score = 24. Further, there was a significant correlation between measures of working memory capacity and RAPM in both conditions (see Table 4). The correlation between RAPM and the composite working memory score was significantly greater by .16 in the given-rules condition, $z(598) = 2.77$, $p < .01$.

Table 3
Task Means, Standard Deviations, and Reliability Estimates for Each Experimental Group in Experiment 1

Tasks		Control	Given-Rules	*d*
SWM (13 items)	*M*	5.80	6.02	0.05
	SD	2.53	2.62	
	n	306	313	
	α	.79	.82	
BDS (12 items)	*M*	6.55	6.47	-0.04
	SD	1.92	2.06	
	n	311	325	
	α	.72	.76	
VS (10 items)	*M*	6.67	6.70	0.03
	SD	1.73	1.84	
	n	311	320	
	α	.79	.80	
WMC (z-score)	*M*	-0.01	0.02	0.04
	SD	0.81	0.85	
	n	303	313	
RAPM (26 items)	*M*	9.60	11.93	0.52
	SD	4.56	4.55	
	n	305	321	
	α	.82	.82	

Note. SWM = spatial working memory, BDS = backward digit span, VS = verbal span, WMC = working memory composite score, RAPM = Raven Advanced Progressive Matrices.

Table 4
Correlations Among Different Tasks for Each Experimental Group in Experiment 1

	WMC	SWM	BDS	VS
Control Condition:				
RAPM	.46 (292)	.42 (295)	.33 (300)	.31 (300)
WMC		.84 (303)	.83 (303)	.46 (301)
SWM			.39 (303)	.27 (303)
BDS				.49 (308)
Given-Rules Condition:				
RAPM	.62* (308)	.58* (308)	.46* (318)	.38 (313)
WMC		.84 (313)	.84 (313)	.46 (307)
SWM			.41 (313)	.31 (307)
BDS				.49 (318)

Note. *r* (*N*). SWM = spatial working memory, BDS = backward digit span, VS = verbal span, WMC = working memory composite score, RAPM = Raven Advanced Progressive Matrices. * indicates significant difference from control group at *p* < .05

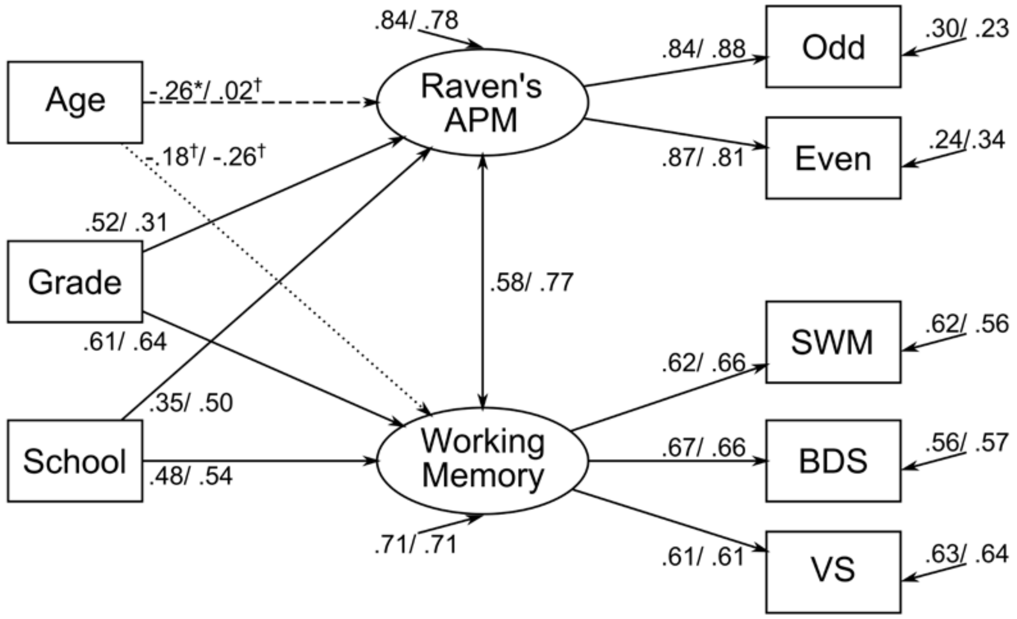


Figure 2. Experiment 1 Model 1. Standardized estimates for control/ given-rules condition. SWM = spatial working memory, BDS = backward digit span, VS = verbal span. $^{\dagger}p = ns$, $^{*}p < .05$, all other estimates are significant at $p < .01$.

Third, a latent variable model was estimated in order to account for measurement error and possible covariates. All calculations were performed using the software Mplus 6.1. In a first step, a standard model was estimated (with maximum likelihood) for the two experimental groups. A latent factor for performance on RAPM was created using the sum scores of odd and even test items. Both factor loadings were fixed to 1 since both parts are expected to represent the underlying factor equally. Working memory capacity was modeled by performance on spatial working memory, backward digit span, and verbal span. All factor loadings were restricted to be equal across groups to specify metric invariance. Also, the participants' age, grade, and school level were included in the analysis as covariates. The resulting parameter estimates are depicted in Figure 2 and the model fitted reasonably well ($\chi^2 = 73.52$, $df = 33$, $p < .01$, RMSEA = .06, CFI = .97, TLI = .95). Estimates for the correlation between RAPM and WM were $r = .58$ in the control group and $r = .77$ in the given-rules group, which suggests an increase of .19 due to the experimental manipulation. In order to see if this difference is significant, a restricted model was estimated, in which the correlation between RAPM and WM was constrained to be equal across groups. The resulting model also fitted reasonably well (χ^2 = 78.23, *df* = 34, *p* < .01, RMSEA = .06, CFI = .96, TLI = .94), yet a chi-square difference test indicated that the fit of this restricted model was significantly worse than the fit of the standard model ($\Delta\chi^2$ = 4.71, *df* = 1, *p* = .03), suggesting the correlation should not be restricted to be equal across groups. That is, the difference of the correlation between groups was significantly different from zero.

Discussion

Based on the theoretical assumption that the solution process in RAPM can be divided into two subprocesses, rule induction and goal management (Carpenter et al., 1990), we explored the extent to which goal management processes might explain the relation between working memory capacity and RAPM performance in a condition where rule induction might not be required. To test this, Experiment 1 incorporated a manipulation of RAPM in which the appropriate rules for solution had been taught to the participants beforehand. Results revealed that the correlation between RAPM and measures of working memory did not decrease when participants were given the rules, suggesting that goal management is correlated with working memory capacity. Further analyses revealed that the correlation significantly increased due to the experimental manipulation. Although the overall solution rates are higher when the rules had been given to participants, the relative account of working memory for task performance was larger. The difference of the correlations between the two experimental conditions was evident from zero-order correlations at single task level, as well as from first-order latent variable correlations. A possible interpretation of the present results is that teaching participants the rules eliminates the need for rule induction, and makes the task more clearly one of goal management, which itself is highly related to working memory capacity.

Experiment 2

The purpose of this experiment was to further investigate what happens when the rules are known while solving RAPM. In Experiment 1, it was shown that teaching participants the rules to solve RAPM problems did not decrease, but did in fact increase the relation between working memory capacity and RAPM. It was suggested that teaching participants the rules removed the need to engage in rule induction. The aim for this experiment was to operationalize and measure rule induction ability as a way of showing that this ability might be less predictive of success on RAPM when the rules are given. Thus, if rule induction is not necessary any more, RAPM should not correlate with measures of rule induction ability, or any correlation with a measure of rule induction ability should significantly decrease under the given-rules condition.

Method

Measuring Rule Induction Ability

Rule induction ability was operationalized via the Brixton spatial rule anticipation test (BRX), as described by Crescentini et al. (2011). This test requires predicting the pattern of change in a visuo-spatial stimulus. In the original version, the stimulus changes according to a hidden rule and the rule might change without notice (similar to the Wisconsin Card Sorting Test). The test demands the frequent induction of new rules according to the most recent sequence of stimuli, and the inhibition of previously learned rules. Some changes were made compared to Crescentini et al. (2011) in order to make the test more suitable for the targeted age group and to focus more on rule induction than on rule inhibition. To reach the former goal, 10 rule sequences from the lower end of the difficulty spectrum were selected. To achieve the latter goal, the test was altered in a way that rules would not change unannounced. Instead, participants were presented 10 distinct items with a new rule on each.

On every item, participants were presented with a 2x6 arrangement of empty circles. One of these circles would fill with blue color for one second. After that, participants were asked to predict which circle would fill next. After participants indicated their response the next circle would fill with blue color for one second and the participants received a feedback whether their prediction was correct or not. After a certain amount of steps, the visuo-spatial patterns became predictable in that they followed a certain rule. Once participants detected the hidden rule, they should have been able to predict the next step.

The test consisted of 10 items plus two practice items. Rule patterns could vary in regard to their period, meaning the number of steps before the rule starts repeating and becomes predictable. Items with period two consisted of 11 steps, and items with period one consisted of eight steps. Once a participant correctly predicted at least two consecutive steps without further errors, the rule was considered detected. If participants failed to predict a rule they were assigned a score equal to the steps in that item plus one. The mean step at which a correct sequence begins across all 10 items served as the dependent variable and as an indicator of rule induction ability. Thus, lower scores on this task indicate better rule induction ability.

Working Memory Assessment

The same complex span measures as in Experiment 1 were used, excluding the verbal span task. Previous studies suggested that the intelligence-working memory correlation may be accounted for by short term storage, presumably because all of these constructs are limited by the same cognitive resources to a very high degree (Colom, Abad, Quiroga, Shih, & Flores-Mendoza, 2008; Martínez et al., 2011). Accordingly, two measures of short term memory capacity were added to assess the unique influences of working memory capacity over and above short term memory capacity on RAPM performance.

The first measure was a forward digit span task (FDS). This measure was comparable to the backward digit span used in Experiment 1, but participants were required to recall the digits in the same order as they were presented. This task consisted of two practice items and 10 test items, whereas the length of series was increasing from 4 to 8. Applying partial credit scoring, the maximum possible score was 10.

The other measure was a dot memory task (DM) which was adapted from Miyake, Friedman, Rettinger, Shah, and Hegarty (2001). This task required memorizing a visuo-spatial pattern over a short time interval and then recalling that pattern. Participants were presented with a 5x5 matrix of white squares on grey foreground. A varying number of blue dots appeared simultaneously in a subset of these squares for one second, then the matrix was masked black for 50ms and finally all squares where white again. The number of dots in a stimulus was increasing from 3 to 7 across items. The participants were then asked to indicate their memory of the locations of the blue dots by clicking on the corresponding white squares. There was a 60 second time restriction on the response screen. After that or when the participants pressed an ok-button, the matrix was masked for 50ms again, and then the next item was presented. Two practice items preceded 10 test items. Using partial credit scoring the maximum possible score was 10.

Two composite measures were created, one for working memory capacity and one for short term memory capacity, each by averaging z-scores of the corresponding measures for each construct.

Procedure

The participants were tested in small groups not larger than 10 individuals. All tasks were presented on tablet PCs with touchscreen. Participants worked on the working memory tasks first, followed by the Brixton test, then RAPM, and finally the short term memory tasks. Other than that, the procedure was the same as in Experiment 1.

Sample

Again, secondary school students were recruited from the same population and with the same procedure as in Experiment 1, but from 4 different schools. The total sample comprises of *N* = 366 individuals, with *n* = 176 in the control condition and *n* = 190 in the given-rules condition. There were 158 students from grade 5, 54 from grade 6, 118 from grade 7, and 36 from grade 8. There were 196 students from the highest school level and n = 170 students from the second highest school level (only Gesamtschule in this sample). About 51% of the participants were female. The participants age ranged from 10 to 16 years (*M* = 12.1, *SD* = 1.2).

Results

First, differences between the experimental conditions in working memory capacity, short term memory capacity, and the Brixton test were examined (see Table 5). Unfortunately, random assignment did not result in equal distributions between the two conditions. Significant differences were present between groups on spatial working memory, *t*(361) = 2.00, p = .05 and on dot memory *t*(320.19) = 3.08, *p* < .01. Levene's test of homogeneity of variances was significant for dot memory meaning that the two conditions differed in their variance on this task, *F*(1, 356) = 6.35, *p* = .01. And some differences were seen in patterns of relations among the working memory and short term memory measures (see Table 6), namely a higher correlation between dot memory and forward digit span in the given-rules condition, *z*(357) = 2.66, *p* < .01, and the correlation between dot memory and backward digit span was slightly higher in the given-rules condition, however not significantly, *z*(353) = 1.79, *p* = .07. There was no significant difference on RAPM score between groups, *t*(338.17) = 1.25, *p* = .21. Furthermore, Levene's test of homogeneity of variances was significant for RAPM, *F*(1, 351) = 4.87, *p* = .03.

Table 5

Task Means, Standard Deviations, and Reliability Estimates for Each Experimental Group in Experiment 2

Tasks		Control	Given-Rules	*d*
SWM (13 items)	*M*	5.38	4.71	-0.21
	SD	3.29	3.09	
	n	174	189	
	α	.88	.86	
BDS (12 items)	*M*	6.20	5.89	-0.13
	SD	2.42	2.32	
	n	173	189	
	α	.84	.81	
FDS (10 items)	*M*	6.00	5.84	-0.09
	SD	1.73	1.97	
	n	175	188	
	α	.77	.81	
DM (10 items)	*M*	8.54	8.12	-0.33
	SD	1.01	1.54	
	n	173	185	
	α	.74	.86	
WMC (z-score)	*M*	0.09	-0.08	-0.20
	SD	0.88	0.85	
	n	173	189	
STMC (z-score)	*M*	0.11	-0.10	-0.25
	SD	0.68	0.97	
	n	173	185	
BRX (35 items)	*M*	6.33	6.42	0.05
	SD	2.04	1.83	
	n	174	187	
	α	.80	.74	
RAPM (26 items)	*M*	10.05	10.74	0.14
	SD	5.42	4.80	
	n	170	183	
	α	.87	.83	

Note. SWM = spatial working memory, BDS = backward digit span, FDS = forward digit span, DM = dot memory, WMC = working memory composite, STMC = short term memory composite, BRX = Brixton test, RAPM = Raven Advanced Progressive Matrices.

Table 6
Correlations Among Different Tasks for Each Experimental Group in Experiment 2

	BRX	WMC	STMC	SWM	BDS	FDS	DM
Control Condition:							
RAPM	-.57 (169)	.68 (169)	.52 (167)	.69 (169)	.49 (168)	.43 (169)	.39 (167)
BRX		-.59 (172)	-.50 (171)	-.56 (173)	-.44 (172)	-.37 (173)	-.43 (171)
WMC			.64 (170)	.86 (174)	.86 (173)	.60 (172)	.39 (170)
STMC				.50 (171)	.59 (170)	.84 (173)	.75 (173)
SWM					.47 (173)	.39 (173)	.41 (171)
BDS						.65 (172)	.25 (170)
FDS							.26 (173)
Given-Rules Condition:							
RAPM	-.57 (182)	.68 (182)	.52 (180)	.64 (182)	.55 (182)	.49 (183)	.39 (180)
BRX		-.57 (187)	-.43 (184)	-.55 (187)	-.44 (187)	-.39 (187)	-.35 (184)
WMC			.61 (184)	.87 (189)	.88 (189)	.62 (187)	.44 (184)
STMC				.47 (184)	.59 (184)	.86 (185)	.88 (185)
SWM					.53 (189)	.48 (187)	.35 (184)
BDS						.61 (187)	.42 (184)
FDS							.50* (185)

Note. *r* (*N*). SWM = spatial working memory, BDS = backward digit span, FDS = forward digit span, DM = dot memory, RAPM = Raven Advanced Progressive Matrices, BRX = Brixton test, WMC = working memory composite, STMC = short term memory composite. * indicates significant difference from control group at *p* < .05

As shown in Table 6, correlations were seen between RAPM and working memory composite score, and RAPM and the Brixton test, in both conditions. There was no difference in the strength of these correlations between conditions.

Second, because random assignment failed, a subsample of participants was matched on the working memory composite score. The matching procedure searched for each individual in the control group another individual in the experimental group with a comparable value on the working memory composite raw score. A comparable value was defined as having a maximum difference of +1 or -1. In terms of z-scores this means that values differed no more than +0.5 or -0.5. This resulted in a sample of 320 participants, 160 in each condition. This matched sample was used for further analyses. In the matched sample, participants in the given-rules condition solved more problems than participants in the control condition, although in this experiment the advantage was only about 1.5 problems, *t*(308) = 2.70, *p* < .01. Still, there were no group differences in correlations. To test for the unique effects of each predictor variable, a linear regression with RAPM as the dependent variable showed that working memory composite, *β* = .43, *t*(299) = 7.54, *p* < .01, and the Brixton test, *β* = -.29, *t*(299)

= 5.72, $p < .01$, both contributed significant unique variance when short term memory composite, $\beta = .10$, $t(299) = 1.99$, $p = .05$, was included in the model, $R^2 = .49$, $F(3, 299) = 94.26$, $p < .01$.

Fitting a comparable model as in Experiment 1 to the data with one factor for RAPM and one for working memory capacity, indicated by spatial working memory and backward digit span, revealed that the factors were correlated at .91 in the control condition and at .90 in the given-rules condition ($\chi^2 = 8.62$, $df = 8$, $p = .38$, RMSEA = .02, CFI = .99, TLI = .99). Adding a short term memory factor, indicated by forward digit span and dot memory, did not produce a well-fitting model ($\chi^2 = 92.59$, $df = 21$, $p < .01$, RMSEA = .14, CFI = .93, TLI = .89) and resulted in an unrealistic estimate for the correlation between short term memory and working memory in the control condition of 1.04, indicating a possible linear dependency between factors. Thus, a further model was fitted where the working memory factor was indicated by all four memory tasks ($\chi^2 = 104.43$, $df = 26$, $p < .01$, RMSEA = .13, CFI = .92, TLI = .91) and estimated the correlation between working memory and RAPM at .86 in the control condition and .87 in the given-rules condition. Since previous analyses suggested that dot memory and spatial working memory differed between the two groups, we then excluded those tasks and fitted a two factor model with working memory being indicated by backward digit span and forward digit span only. This model fitted quite well ($\chi^2 = 3.84$, $df = 6$, $p = .70$, RMSEA < .01, CFI = 1, TLI = 1) and estimated the correlation between factors at .61 in the control condition and at .74 in the experimental condition, which is comparable to the estimates obtained in Experiment 1. However, restricting the factor correlation to be equal between groups did not result in a significantly worse fitting model ($\Delta\chi^2 = 2.32$, $df = 1$, $p = .13$), indicating this difference is not statistically significant. We finally added the Brixton score as an observed variable to the unrestricted model regressed on the working memory factor while also estimating correlations with RAPM factor ($\chi^2 = 5.16$, $df = 10$, $p = .88$, RMSEA < .01, CFI = 1, TLI = 1). The resulting estimates for the correlation between RAPM and the Brixton test was .31 in the control group and .23 in the given-rules group. The beta weight between the working memory factor and Brixton score was estimated at .43 in the control group and at .52 in the given-rules group. None of these differences qualified to be statistically significant.

Discussion

The purpose of Experiment 2 was to replicate and extend the findings of Experiment 1. We predicted that working memory capacity would again explain a larger amount of unique variance in performance on RAPM when rules were given to participants. Second, we predicted that rule induction ability would not correlate with RAPM in the given-rules condition.

Contrary to predictions, the results of this experiment do not indicate a difference in the correlation between RAPM and working memory capacity between groups. Comparing these correlations across studies, it seems like the correlation in the control group was unusually high, compared to Experiment 1 and also compared to values obtained in previous research (eg., Ackerman et al., 2005). The Brixton test was highly correlated with RAPM in both conditions and the correlation did not differ as predicted.

The samples in Experiment 1 and 2 were drawn from the same population and were comparable in terms of age and grade. In light of these similarities it is surprising that the variances in Experiment 2 were unequal and that some measures of working memory capacity and short term memory suggested differences between the two conditions. Only when these differences were reduced by using a matched sample, a small benefit could be seen from being given the rules in this study. Furthermore, excluding the odd tasks from the structural equation model revealed that results were much more in line with results obtained in Experiment 1. Comparing the levels of performance across the two studies, it appears that the given-rules group did less well in Experiment 2, indicating that the experimental manipulation might not have had the same impact.

Some experimental circumstances may be noteworthy at this point. First, in some schools testing conditions may have been suboptimal. Some of the rooms where the testing took place were rather tiny and participants were sitting fairly close to each other. This resulted in a lot of interaction among students. Some of them might have peeked, hence blurring experimental differences. Another difference from the procedure in Study 1 was the addition of the Brixton test before RAPM. This rule induction test was frustrating for many students since mistakes are necessary, and this test being in advance to the experimental manipulation might have had an effect on it. Additionally the number of tasks was increased from 4 to 6 compared to Experiment 1 which some students remarked as being too much work.

These imperfect experimental circumstances might give some reasons for the failed manipulation in this experiment from the standpoint of affecting the relation between working memory capacity and RAPM. We must, however, acknowledge the possibility that the results obtained in Experiment 1 might just have been a false positive.

Experiment 3

The primary purpose of Experiment 3 was to test whether a measure of productive thinking might differentially predict performance across the given-rules and not-given-rules conditions. One way to think about rule induction is that it requires the ability to come up with new approaches to a problem and generate new rules. A construct that seems to be related to this idea is productive thinking which in turn is tapped by divergent thinking tasks, such as thinking of novel uses for a brick (Guilford, 1957). Generally measures of this kind are not very highly correlated with general intelligence measures (K. H. Kim, 2005), but based on findings from Nusbaum and Silvia (2011) we expected to find something in the range of .10 to .20. Assuming that divergent thinking is not necessary under the given-rules condition, we expected this correlation to drop near zero.

A second purpose of Experiment 3 was to see if the findings of Experiment 1 could be replicated under more controlled experimental circumstances. As such we predicted to find that measures of working memory would be strongly correlated with RAPM in the given-rules condition, and less so in the control condition.

Method

Measuring Productive Thinking

A very frequently used task to measure productive thinking is the unusual uses task, where one has to come up with unusual ideas for use of everyday objects (Silvia et al., 2008). For this experiment a computerized adaptation of the task was used. Prior to the task, participants watched a video where they learned that the task was about creativity and that it was important to generate responses that would be as creative as possible. Creative responses were defined as being original, unusual, and sometimes funny. An example item was given (unusual uses of duct tape) along with some example responses. Participants were asked to generate unusual uses for three items: a brick, a wooden board, and a paper coffee cup. They were given two minutes time for each item indicated by a timer counting down backwards. On the left-hand side of the screen was a picture of the object along with a prompt asking what the object could be used for. On the right-hand side of the screen was a textbox where participants could type in their responses. After the time has expired the participants were asked to indicate which of their responses were the most creative by clicking a checkbox next to the response.

The responses were later independently coded by six raters (research assistants) on a Likert scale ranging from 1 (not at all creative) to 5 (very creative). Individual responses were pooled in one table and the order randomized so that raters were blind to condition and other characteristics of respondents. In order to ensure raters had the same conception of creativity as the respondents they were told that responses, in order to qualify as creative, should be original, unusual, or funny. There was reasonable agreement among raters for ratings of unique responses (see Table 7). Only ratings for the two most creative responses, indicated by participants themselves, were considered for the total score (top two scoring method, Silvia et al., 2008). These ratings were averaged across all six raters to obtain one creativity score on each item. Fluency, or the number of valid responses produced, was used as an additional indicator of productive thinking.

Table 7

Correlations Across Creativity Ratings of Unique Responses for Each Productive Thinking Item in Experiment 3

Item		Rater 2	Rater 3	Rater 4	Rater 5	Rater 6
Brick	Rater 1	.20	.49	.27	.13	.38
	Rater 2		.34	.49	.58	.33
	Rater 3			.60	.29	.42
	Rater 4				.50	.41
	Rater 5					.28
Board	Rater 1	.10	.29	-.01	.09	.25
	Rater 2		.29	.54	.58	.34
	Rater 3			.34	.26	.45
	Rater 4				.55	.41
	Rater 5					.30
Cup	Rater 1	.19	.28	.08	.09	.37
	Rater 2		.36	.41	.63	.34
	Rater 3			.60	.39	.46
	Rater 4				.47	.40
	Rater 5					.34

Note. The total number of unique responses (*N*) was 308, 332, and 248 for the Brick, Board, and Cup item respectively.

Working Memory Assessment

To measure working memory capacity, the same complex span tasks as in Experiment 2 were used (spatial working memory and backward digit span) plus reading span as an additional verbal task. Like the other working memory measures, this task was adapted from Vock and Holling (2008). The reading span task (RS) presents a series of written statements on the screen which were either logical (e.g. a snowman is made of snow) or nonsensical (e.g. humans have three legs). Each statement was displayed for 5 seconds during which the participants had time to give a response about whether that statement was true or false. Participants were instructed to memorize the final word of each statement for later recall. The series varied in length between 3 and 6. After the series, a prompt asked to enter the memorized words in a textbox. There was a 60 second time restriction on the response screen. Three practice items preceded the 11 test items. Each correctly recalled word being in the correct order was scored with one point divided by the number of words on the item (partial credit scoring), for a maximum possible score of 11 points. Errors of commission and errors of omission were ignored. For consistency across studies, a composite working memory score based on z-scores of spatial working memory and backward digit span is used for the main analysis (including reading span did not affect the pattern of results).

Raven's APM

The same reduced version of RAPM as in Experiments 1 and 2 was used, but this time with a manipulation check added. Right after the practice set, participants in the given-rules condition were prompted to recall what rules they had learned in the instruction video (free recall test). After entering a response they saw the list of correct responses as feedback. This served not only as a manipulation check, but also as a form of reinforcement by retrieval-based learning. After that, participants would see a depiction of a rule from the introduction video and were given 5 multiple choice response options, one for each rule, to indicate which rule was depicted (recognition test). This was repeated for two more items. We later realized that the depictions were ambiguous, meaning that for some of them, more than one rule was applicable, therefore this part of the manipulation check will be disregarded in future analyses (however, the results were comparable to the free recall test).

Procedure

In this study a couple of measures were taken to ensure a more controlled environment during testing sessions. It was ensured that participants had enough space and dividers were placed in between them, so they would not peek or interact with each other during testing. Feedback was added to the working memory tasks, so after each item the students were shown how their response contrasted against the correct response. Many students have asked about this in our previous studies, consequently this might have improved compliance and prevented frustration. Additionally, the introduction videos for all tasks were improved in audio and video quality, but with the same text and video content. Furthermore, the unusual uses task was put at the end, after RAPM, so the task order was comparable to Experiment 1. Other than that, the procedure was the same as in Experiment 1.

Sample

Again, secondary school students were recruited from the same population and with the same procedure as in Experiments 1 and 2, but from 6 different schools. The total sample consisted of N = 393 individuals, with n = 199 in the control condition and *n* = 194 in the given-rules condition. There were 79 students from grade 5, 185 from grade 6, and 129 from grade 7. There were n = 52 individuals from the highest school level and 341 individuals from the second highest school level (Realschule and Gesamtschule combined). About 53% of the participants were female. The participants age ranged from 10 to 14 years (*M* = 11.8, *SD* = 0.9).

Results

First, means and differences in task performance are reported for the two experimental groups, which can be obtained from Table 8. Task performance in the working memory tasks did not differ significantly between groups (*t*-values from 0.96 to 1.31, *df* from 373 to 380, *p*-values from .19 to .54) and the variances of the working memory tasks did not differ significantly between groups (Levene's *F*-values from 0.03 to 2.54, *p*-values from .11 to .87). Also, the correlations among the three working memory tasks did not differ significantly between experimental groups (see Table 9). This suggests that random assignment resulted in an evenly distributed working memory capacity profile in both experimental groups. This also holds for measures of productive thinking, except the correlation between fluency and creativity ratings for the board item was significantly lower in the given-rules condition, *z*(363) = 1.97, *p* = .05 .

Table 8
Task Means, Standard Deviations, and Reliability Estimates for Each Experimental Group in Experiment 3

		Control	Given-Rules	*d*
SWM (13 items)	*M*	4.62	4.26	-0.16
	SD	2.53	2.75	
	n	192	190	
	α	.78	.84	
BDS (12 items)	*M*	5.54	5.66	0.06
	SD	1.86	1.96	
	n	195	191	
	α	.73	.75	
RS (11 items)	*M*	6.76	6.57	-0.10
	SD	1.78	1.98	
	n	188	187	
	α	.85	.86	
WMC (z-scores)	*M*	,01	-,01	-0.02
	SD	,82	,86	
	n	195	192	
Average Fluency (3 items)	*M*	4.29	4.33	0.02
	SD	1.76	1.52	
	n	195	190	
	α	.80	.75	
Brick Average Creativity	*M*	2.26	2.24	-0,04
	SD	0.57	0.51	
	n	189	185	
Board Average Creativity	*M*	2.40	2.42	0,04
	SD	0.46	0.46	
	n	187	178	
Cup Average Creativity	*M*	2.26	2.32	0,12
	SD	0.49	0.52	
	n	187	181	
RAPM (26 items)	*M*	6.61	8.89	0.56
	SD	4.20	3.93	
	n	191	191	
	α	.79	.75	

Note. SWM = spatial working memory, BDS = backward digit span, RS = reading span, RAPM = Raven Advanced Progressive Matrices.

Second, participants in the given-rules condition solved about 2.3 items more due to the experimental manipulation, *t*(380) = 5.47, *p* < .01, *d* = .56. Furthermore, the manipulation check revealed that participants in the given-rules condition recalled on average 2.8 out of 5 rules (*SD* = 1.6) on the free recall test. About 81% of participants recalled at least one rule, and about 70% recalled at least 3 rules, indicating that the manipulation was successful for the majority of participants. The dark shaded bar in Figure 3 marks the mean RAPM score in the control group, which is almost exactly as high as the score of participants who did not recall any rules in the given-rules group, while subjects who recalled at least one rule scored higher on RAPM. The amount of rules recalled was significantly correlated with RAPM, *r*(186) = .40, *p* < .01, and this correlation remains significant after controlling for working memory capacity, *r*(183) = .23, *p* < .01. This indicates that knowing the rules does help in solving RAPM items.

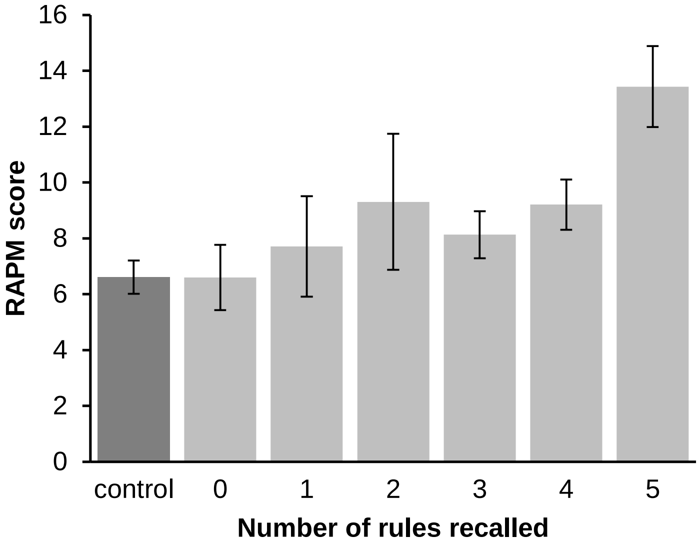


Figure 3. Mean RAPM scores in Experiment 3 depending on the number of rules recalled in the given-rules condition. The dark shaded bar represents the mean in the control condition. Error bars represent 95% CI.

Third, an inspection of Table 9 reveals that the correlation between RAPM and working memory composite was significantly greater by .19 in the given-rules condition, *z*(376) = 2.55, *p* = .01. Note, that the magnitude of the correlations was comparable to the ones observed in Experiment 1.

Fitting a comparable model as in Experiment 1 (χ^2 = 44.87, *df* = 27, *p* = .02, RMSEA = .06, CFI = .97, TLI = .96) suggests that the correlation between RAPM and working memory capacity in the given-rules condition is about .29 higher than in the control condition (see Figure 4). Again, restricting this difference to be zero resulted in a significant decrease in model fit ($\Delta\chi^2 = 7.10$, $df = 1$, $p < .01$), suggesting the difference is significantly different from zero.

Finally, there were no significant differences between groups in the correlations between RAPM and measures of productive thinking (see Table 9). Exploring the relation in a structural equation model with productive thinking operationalized as fluency ($\chi^2 = 65.46$, $df = 45$, $p = .02$, RMSEA = .05, CFI = .98, TLI = .97) revealed that fluency is not significantly correlated with RAPM in either condition (see Figure 5). The same was true when operationalizing productive thinking via creativity ratings ($\chi^2 = 61.37$, $df = 45$, $p = .05$, RMSEA = .04, CFI = .98, TLI = .97; see Figure 6).

Table 9
Correlations Among Different Tasks for Each Experimental Group in Experiment 3

	Fluency	Brick	Board	Cup	SWM	BDS	RS	WMC
Control Condition:								
RAPM	.14 (190)	.15 (185)	.14 (182)	.11 (182)	.36 (187)	.33 (189)	.39 (183)	.42 (189)
Fluency		.27 (189)	.29 (187)	.13 (187)	.13 (190)	.14 (192)	.33 (186)	.15 (192)
Brick			.33 (181)	.33 (182)	.08 (184)	.14 (186)	.17 (180)	.13 (186)
Board				.27 (183)	.18 (182)	.17 (184)	.22 (178)	.20 (184)
Cup					.08 (182)	.09 (184)	.04 (178)	.10 (184)
SWM						.44 (192)	.36 (185)	.84 (192)
BDS							.57 (187)	.85 (195)
RS								.54 (187)
Given-Rules Condition:								
RAPM	.17 (189)	.23 (184)	.31 (177)	.23 (180)	.58* (187)	.44 (188)	.50 (184)	.61* (189)
Fluency		.15 (185)	.09* (178)	.13 (181)	.19 (187)	.18 (188)	.28 (183)	.21 (189)
Brick			.27 (173)	.22 (176)	.24 (182)	.13 (183)	.11 (178)	.23 (184)
Board				.18 (174)	.20 (175)	.12 (176)	.09 (171)	.20 (177)
Cup					.17 (179)	.19 (179)	.23 (174)	.23 (180)
SWM						.36 (189)	.37 (183)	.83 (190)
BDS							.53 (184)	.83 (191)
RS								.54 (185)

Note. *r* (*N*). SWM = spatial working memory, BDS = backward digit span, RS = reading span, RAPM = Raven Advanced Progressive Matrices, WMC = working memory composite. * indicates significant difference from control group at $p < .05$.

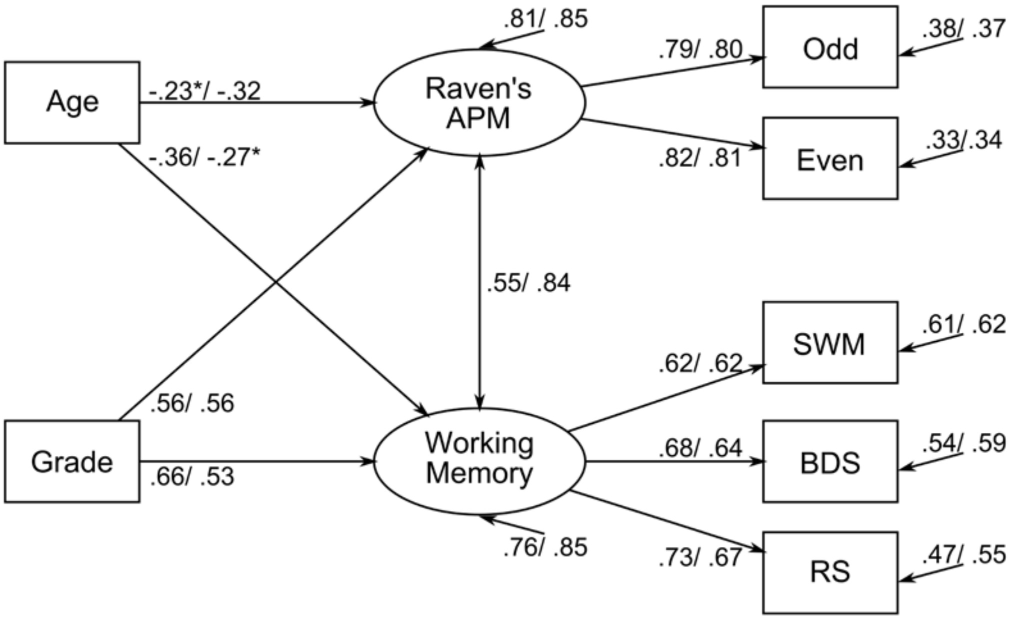


Figure 4. Experiment 3 Model 1. Standardized estimates for control/ given-rules condition. SWM = spatial working memory, BDS = backward digit span, RS = reading span. **p* < .05, all other estimates are significant at *p* < .01.

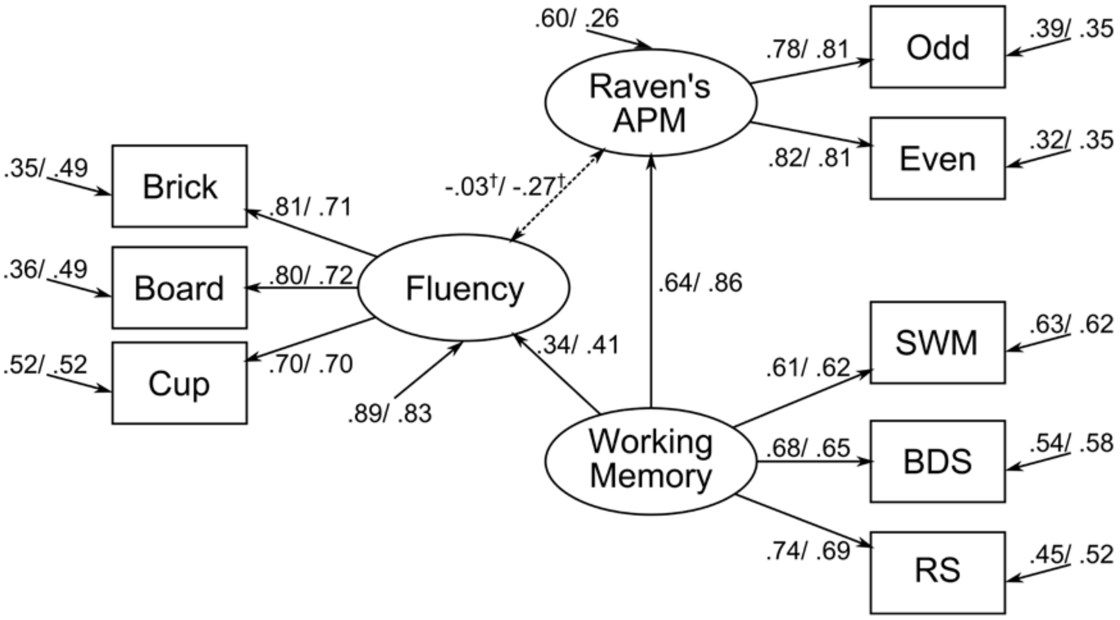


Figure 5. Experiment 3 Model 2. Standardized estimates for control/ given-rules condition. SWM = spatial working memory, BDS = backward digit span, RS = reading span. †*p* = *ns*, all other estimates are significant at *p* < .01.

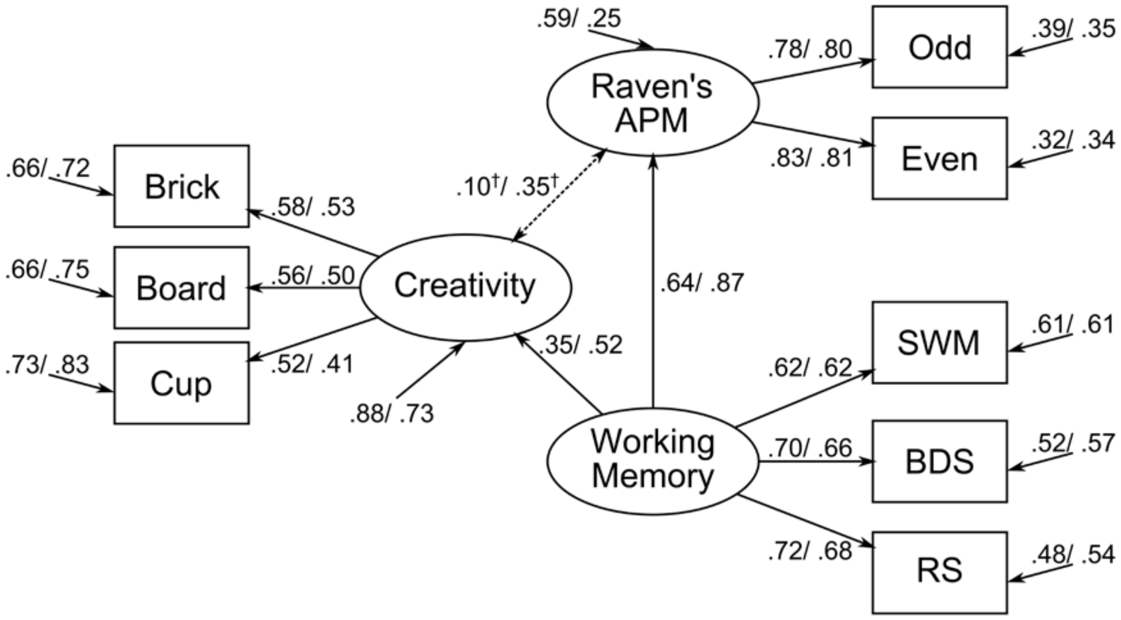


Figure 6. Experiment 3 Model 3. Standardized estimates for control/ given-rules condition. SWM = spatial working memory, BDS = backward digit span, RS = reading span. $^{\dagger}p = ns$, all other estimates are significant at $p < .01$.

Discussion

The results from Experiment 3 are much more in line with Experiment 1, replicating the original finding, that RAPM is more highly correlated with measures of working memory capacity when the rules are known. Thus, the results suggest that this relationship is even stronger when the task is reduced to goal management. The findings were extended and corroborated by a manipulation check, which indicated that subjects fair better on RAPM when they have learned more rules. More importantly, the results of the manipulation check revealed that the majority of the subsample retained at least some of the rules. We further explored whether RAPM would correlate less with a measure of productive thinking when the rules were given, which was not the case. After accounting for working memory capacity operationalized via fluency or via creativity ratings RAPM was not correlated at all with productive thinking in either condition.

Experiment 4

The previous experiments have shown twice that the relation between working memory capacity and solution on a subset of RAPM problems increases when participants are taught the rules required for solving those problems. However, it is still an open question how providing the rules is changing the task, and why it is increasing the relationship between working memory capacity and RAPM. Two attempts failed to demonstrate that reliance on rule induction or productive thinking processes is reduced in those conditions. Measures of rule induction and productive thinking were found to be just as correlated with performance in the given-rules conditions as in the no-rules condition. Thus, a new approach to understanding the effect was undertaken in this final study. In this study, we used eye tracking measures to test whether knowledge of the rules might be affecting the strategies that students deploy while attempting to solve RAPM problems.

Bethell-Fox, Lohman, and Snow (1984) described two different strategies that solvers appear to use on figural analogy problems: constructive matching and response elimination. They suggested that constructive matching would be predominantly applied on easy problems and by high ability subjects. This result was corroborated and further extended by Vigneau, Caissie, and Bors (2006) in an eye tracking study with Raven's APM. They found that subjects who scored higher on Raven's APM spent proportionately more time analyzing the actual problem and waited later before consulting response alternatives. They interpreted this behavior as reflecting the reliance on matrix information as opposed to information from the response alternatives, which is consistent with the idea of a constructive matching strategy.

Coming back to Carpenter et al. (1990), the way they described the solution process very much resembles the constructive matching strategy. This is not surprising, given that their simulation models were informed by eye movement patterns of an elite university student sample. Their computer programs compared elements in the matrix row wise one by one to extract differences and similarities. These comparisons used the build-in list of rules as a reference to see if any of those could account for observed regularities. This process was called correspondence finding, and it continued until all unaccounted elements were covered. Based on that, a further process generated preliminary responses and compared them with response alternatives. Hence, the computer programs simulated a constructive matching strategy under the condition of known rules.

Thus, Carpenter et al.'s computer models provide a prediction for the eye movement behavior we may see in real humans while solving RAPM in the given-rules condition. In the given rules condition, it can be hypothesized that eye movements should be more consistent with a constructive matching strategy. Subjects should spend more time attempting to identify matching elements to rules in order to come up with a possible response. That way, they should only later consult response alternatives for further information. On the other hand, in the control condition where rules are unknown, we hypothesized that subjects would earlier and more frequently seek additional information from the response alternatives. Since subjects have no framework of rules to guide their correspondence finding behavior, they might consider more sources of information to make sense of the problems.

In order to test these hypotheses, the current experiment combined the same rule teaching paradigm from the previous experiments with eye tracking methodology. Based on Vigneau et al. (2006) we extracted three variables from eye tracking protocols to reflect constructive matching behavior: proportion of time before first fixation on response bank, toggle rate between matrix and response bank, and proportion of total time on matrix.

The time to first fixation on the response bank is the timestamp of the first fixation falling on the response bank area of interest. Due to the fixation cross before each problem, participants are forced to start their inspection in the lower part of the matrix area. They usually continue to analyze the elements in the matrix to only later seek additional information in the response bank. The proportion of time before the first fixation on the response bank was computed by dividing the timestamp for the first fixation on the response bank by the total solving time. We hypothesized time before the first response bank fixation would be longer in the given-rules condition.

The toggle rate reflects the frequency of saccades made between the matrix and the response bank. The toggle rate was calculated by summing all saccades that both start on the matrix and end on the response bank or vice versa, and dividing the sum by the total solving time in seconds. Hence, toggle rate expresses the number of saccades between matrix and response bank per second of solution time. We hypothesized the toggle rate would be lower in the given-rules condition.

The proportion of time spent on the matrix reflects the overall time of all fixations on the matrix divided by the total time of fixations on either the matrix or the response bank. Fixations outside of the task area are disregarded. We hypothesized that the proportion of time spent on the matrix would be longer in the given-rules condition.

Method

Eye Movement Data Collection

Participants' eye movements during RAPM were recorded using an Eyelink II head-mounted eye tracker. The device samples eye fixations and saccades at a rate of 250Hz with one camera for each eye. A third camera, mounted to the headband, keeps track of infrared signals from the corners of the screen to account for head movement. The eye tracker had 15 min of arc and spatial accuracy of approximately 0.5°. If possible, both eyes were recorded, with only the data from the better-calibrated eye used for analyses. Eye movements were analyzed with respect to two areas of interest: the matrix and the response bank. These areas were defined in a similar way as Vigneau et al. (2006) defined them with the region of interest for the matrix including the 3x3 problem grid, and the region of interest for the response bank including all 8 response options (see Figure 1).

Raven’s APM

Some changes were made to make RAPM compatible with the eye tracker. First, before each problem the participants saw a fixation cross at the center of the screen in order to allow for drift correction and to ensure that every participant started each problem in the same location. Second the time restriction was removed from the test to get eye tracking data on all problems. Furthermore, instead of the entire Set 1 the control group received only half the items (1, 2, 6, 7, 10, and 12) as practice. That way, practice conditions were equalized between the control and known-rules condition and the difference was more pronounced on rule teaching. Again, a manipulation check was performed, similar to that in Experiment 3, by asking participants to remember the 5 rules in a free recall test right after practice.

Working Memory Assessment

Since previous studies suggest that eye movement strategies are dependent on working memory capacity (Jarosz & Wiley, 2012), participants were asked to complete a backward digit span task (backward digit span) before working on RAPM items. The score would serve to ensure that both experimental conditions are comparable in terms of working memory capacity. This task was different from the span tasks in Experiments 1 to 3 in order to account for the expected higher ability in the target sample. Participants encountered a series of digits, individually presented on the screen for 500ms each, increasing in length from 3 to 8 across trials. After each series, participants were required to enter the series backwards in a textbox. The participants could indicate missing digits with an underscore. There was a 15 second time restriction on the response screen. After that or when participants pressed an ok-button, they would receive a feedback by contrasting their response against the correct response. Two practice items preceded 12 test items. Using partial credit scoring the maximum possible score was 12.

Procedure

Participants were tested in single testing sessions lasting between 30 and 45 minutes. After receiving information about the purpose of the experiment they would work on the working memory span task first, without eye tracking. After that, the eye tracker was calibrated. Then they worked on RAPM items with eye tracking.

Sample

Recruited were 47 undergraduates from the subject pool of the University of Illinois at Chicago, of which 23 were randomly assigned to the control condition and 24 to the given-rules condition. An additional eight participants were tested but excluded from the analysis due to low quality eye tracking data and one more participant was excluded due to very low scores. Participants’ age ranged from 18 to 27 years (*M* = 19.3, *SD* = 1.7) with 53% being female. All participants had normal or corrected to normal vision.

Table 10
Task Means and Standard Deviations in Experiment 4

		Control	Given-Rules	*d*
Backward Digit Span (12 items)	M	8.00	7.72	-0.20
	SD	1.39	1.40	
Raven APM (26 items)	M	13.74	16.29	0.81
	SD	3.29	3.01	
First Response Fixation	M	.44	.52	0.89
	SD	.10	.08	
Toggle Rate	M	0.37	0.30	-0.93
	SD	0.09	0.06	
Time on Matrix	M	.80	.81	0.29
	SD	.04	.03	
	n	23	24	

Results

Inspections of backward digit span means reveals that both experimental groups are comparable in their working memory capacity, *t*(45) = 0.67, *p* = .51 (see Table 10). The two experimental groups were also homogeneous in terms of variance in backward digit span and RAPM (Levene's *F*s < 0.60, *p*s > .44).

Overall, participants in the given-rules condition were able to solve about 2.5 items more than in the control condition, *t*(45) = 2.77, *p* < .01, *d* = 0.81. The manipulation check further revealed that participants in the given-rules condition recalled on average 3.4 out of 5 rules (*SD* = 1.5). About 74% of them recalled at least 3 rules. The number of rules recalled was significantly correlated with RAPM, *r*(21) = .49, *p* = .02, and this was also true after controlling for backward digit span, *r*(20) = .47, *p* = .03. This suggests the experimental manipulation had a similar impact as in the previous experiments.

In terms of the eye movement data, significant differences were seen in two measures. As shown in Table 10, the time before the first response bank fixation was significantly longer in the given-rules condition compared to the control group, *t*(45) = 2.94, *p* < .01, *d* = 0.89. The toggle rate was lower in the given-rules condition, *t*(45) = 3.40, *p* < .01, *d* = -0.93. However, there was no significant difference in the proportionate time of fixations on matrix, *t*(45) = 1.12, *p* = .27, *d* = 0.29.

Discussion

The results of this final study show that providing the rules changes the way that solvers approach RAPM problems. Compared to participants in the control condition, participants in the given-rules condition spent longer on the matrix before viewing the response bank, and made fewer saccades back and forth between the matrix and the response bank. These eye tracking data are consistent with the hypothesis that solvers in the given-rules condition are more likely to approach solutions using a constructive-matching strategy. Thus, these results suggest that providing the rules to participants alters the strategies they use to solve RAPM items.

General Discussion

Summary

Over the course of four experiments we manipulated whether participants were provided with a set of rules that related to a subset of RAPM problems, and examined how knowledge of those rules affected RAPM performance. In Experiment 1 we found that giving the rules made test-takers perform better at RAPM items using those rules, and that it elevated the correlation between working memory capacity and performance. In Experiment 2 we tested whether providing the rules eliminated the need to engage in rule induction by introducing a measure of rule induction ability: the Brixton test. We expected to find that this test would not be correlated with RAPM performance when the rules are given to subjects. This was not the case. The Brixton test correlated rather strongly with RAPM performance in both experimental conditions. Additionally, Experiment 2 failed to replicate the finding from Experiment 1 in terms of the correlation between working memory capacity and RAPM performance being increased in the given-rules condition. One possible explanation for this result was that the experimental circumstances were flawed and thus affected the data. All effects of Experiment 1 were replicated in Experiment 3, giving some confidence in the observation of an elevated correlation between RAPM and working memory capacity due to knowing the rules. The results of the manipulation check in Experiment 3 further revealed that subjects actually learned some of the rules and that rule knowledge helped them solving the test, giving some confidence that the observed effects are due to the knowledge of rules. However, the addition of a productive thinking measure was again unable to demonstrate differences between the two experimental conditions. Thus, Experiment 4 investigated whether giving the rules affects the strategies that participants use on RAPM problems by analyzing participants’ eye movements during the test. The results suggested that subjects who were given the rules were more likely to use a constructive matching strategy, while subjects without knowledge of the rule taxonomy seemed to seek more information from the response bank. Overall the results suggest that when the rules are known, the rate of solution increases, solution becomes more correlated with working memory capacity, and more participants attempt to use a constructive matching strategy.

Implications

These studies were motivated by early work that suggested that two processes, goal management and rule induction, are chiefly responsible for successful performance on RAPM items. The intent behind providing solvers with the rules in these studies was to remove the need to engage in rule induction. Consistent with this approach, one explanation for these results is that giving the rules reduce the solution process to one of goal management, hence the correlation with measures of working memory increases. The reason why performance improves with knowledge of the rules is because of the reduction of necessary cognitive processes for solution. Further, the observed changes in eye movement behavior could be considered as symptoms of the changing task demands towards goal management, whereas the goal management process draws on working memory capacity and executive processes. Goal management could benefit from the ability to control interference from previous representations, subgoals, and strategies (Burgess, Gray, Conway, & Braver, 2011). It is not exactly clear in how far storage capacity is relevant (Chuderski & Neecka, 2012), but it is likely that it helps in maintaining representations of possible solutions and intermediary bindings (Oberauer, Süß, Wilhelm, & Sander, 2007). Our results can thus be seen as consistent with the view that working memory capacity may contribute to RAPM performance through its impact on the goal management process.

At the same time, however, the studies presented here were unable to show any of the hypothesized effects for rule induction ability, specifically that rule induction might be less necessary when the rules were known. This might in part be because we failed to select an appropriate measure of this ability, but seeing the overall pattern of results there is an alternative explanation to consider. Especially the finding that giving the rules affects eye movement behavior raises the question whether alterations in behavior can account for the findings. That is, teaching the rules might inspire subjects to use different strategies, which may be more effective, but also more working memory demanding. Hence, the elevated correlation and the higher solution rates in the given rules condition might just be symptoms of the fact that subjects are using strategies that are more demanding. Rule induction may still play a role in the solution process for RAPM items even when the possible set of rules is known, in that participants still need to judge when each rule fits a problem, and will still need to engage in the correspondence finding process of determining which elements can be used to employ a rule.

Another possible account for the results may be that teaching the rules reduces error variance of the task by narrowing the set size of possible strategic approaches. Accordingly, rule induction would not be considered a cognitive process with systematic influence, but rather as a random effect, meaning strategies would be sampled independently of cognitive ability. This would imply that measuring cognitive ability with Raven's APM would be more accurate when giving test-takers the rules. Note however, that estimates of reliability did not differ as a result of the experimental manipulation in neither dataset. Thus, there is no indication of less error variance when the rules are known.

Turley-Ames and Whitfield (2003) have found that low spans profit more from rehearsal strategy training on working memory tasks, and argued that high spans would naturally come up with efficient strategies. Thus, we explored in a separate analysis whether there were similar patterns in our datasets, by comparing the slopes between experimental groups when RAPM is regressed on working memory. We found no significant difference in all datasets indicating that high spans and low spans profit to the same degree from learning the rules, though by trend, high spans profit somewhat more. All datasets are part of the online supplementary material, so interested readers can follow up on our analyses (see Appendix A).

Limitations

One of the limitations may be that the instructions that informed participants of the rules differ not only in terms of content, but also in terms of length and thoroughness, from the instructions given in the control group. The instruction video in the given-rules condition was twice as long and contained twice as many example problems. Hence, the elevated correlation might be driven by participants' better understanding of the task. Although the experimental group saw more example items, the control condition saw three times as many practice items. Nonetheless, future research is desirable that would keep all other aspects of instruction constant between the two conditions, while maintaining the difference in terms of knowledge of the rules themselves.

A major limitation was the failure to include an independent measure of gf in the study to test whether providing participants with the rules fundamentally changes the nature of the test. Teaching the rules might turn the test into a more reliable and valid instrument by reducing noise, bringing out goal management as the main source of variance in the test. On the other hand, if teaching the rules takes the subprocess of rule induction out of the test, it might as well impact its validity, by reducing the correlation with measures of real-life performance, like school achievement. Either way, the fact that the correlation with working memory changes, raises the question how changing the test by giving the rules affects the correlation with other common measures of gf, for example figural analogies, series completion, or mental rotation.

Another limitation was the fact that only a subset of the items from Raven's APM was utilized in the current studies. This raises the question of whether the results, especially the increase in the correlation between working memory capacity and RAPM in the given-rules condition, would generalize to the full RAPM set which contains several items that do not rely on these five rules and require the induction of new and more complex rules (Wiley et al., 2011). Addressing these limitations is a goal for future work so that we can better understand this effect and the extent to which it may generalize.

Conclusions

One of the main contributions of this work is showing the role that working memory capacity plays in RAPM performance via the goal management process. We hope the rule teaching paradigm inspires other researchers to explore how knowing the rules affects the solution process, and to explore how the rule induction process may, or may not, also be a critical link between working memory and measures of gf. Current models of reasoning (and creativity), oftentimes envision individual differences in working memory capacity to play a crucial role (e.g., Hummel & Holyoak, 2003), but individual differences in working memory capacity alone are not sufficient to fully capture the human potential to form entirely new ideas that come seemingly out of nothing (Dartnall, 2002). Especially in the field of artificial intelligence this part of human intelligence seems hard to capture, yet promises to be crucial in order to model machines that go beyond following predetermined algorithms (Boden, 1998). In light of our results from Experiment 4, it seems that the computer programs developed by Carpenter et al. (1990) were already fairly accurate in simulating the solution process under the condition of known rules. What is still needed, as a complement to this work, is to better understand the mechanisms that underlie the solution process when the rules are not known.

Part 3: The Effects of Rule Knowledge on Eye Movements and Response Time in Matrix Reasoning

Publication Note

The following part was first submitted on 3.3.2016 as an original research article and is pending review in the *Journal of Experimental Psychology: Learning, Memory, and Cognition*: Loesche, P. & Hasselhorn, M. (submitted). The Effects of Rule Knowledge on Eye Movements and Response Time in Matrix Reasoning.

Abstract

Given the potential importance of detecting hidden rules during inductive reasoning, we manipulated whether test-takers of a matrix reasoning test know the underlying rules by introducing a short teaching session in advance to the test. We analyzed test performance, response times, and eye movement behavior from a sample of 109 college students in a series of multi-level models. The results suggest that eye movement behavior under the condition of known rules shifted towards a potentially more efficient strategy. That is, longer fixation on the problem space and less relative saccadic frequency between the main areas of interest. Additionally, we identified two groups of eye movement indicators that were distinctly affected by item difficulty and person ability, suggesting that ability and difficulty have a differential impact on strategy use. By putting eye movement indicators in relation to response times, we got a clearer interpretation of the meaning of these variables and they probably reflect the degree to which mental models build up during reasoning. The results have implications for the interpretation of eye movements and response times as these variables provide a window into the cognitive processes involved in figural reasoning.

Introduction

Figural reasoning tasks are among the traditional approaches to assess intelligence and there is accumulating evidence in the literature that there are essentially two kinds of strategies towards solving figural analogies or matrix reasoning tests (Arendasy & Sommer, 2013; Bethell-Fox et al., 1984; Jarosz & Wiley, 2012; Snow, 1980; Vigneau et al., 2006). Those two strategies were mostly derived from the analyses of eye movements and verbal protocols during problem solving. Probably the first to document these strategies was Snow (1980) who labelled them *constructive matching* and *response elimination*. His description of the constructive matching strategy was apparently influenced by theoretical models of problem solving that were prominent at that time. Mulholland, Pellegrino, and Glaser (1980) compared and reviewed some of these models and noted that the solution process is usually characterized by an iterative series of steps that are similar in all of them. The process usually starts with an encoding phase which results in the identification of visual elements and members of the problem space. Since it is essential to most measures of mental ability, the next step is usually related to finding the underlying rules of the elements in the problem space. The final steps involve the mapping of elements and rules and ideally results in a mental representation of a response and the ultimate choice of a response. The constructive matching strategy basically describes exactly these steps of problem solving and is mostly characterized as effective but also as mentally demanding. Snow (1980) also noticed in his data that occasionally subjects seem to deviate from that strategy as evident from eye movement patterns that showed a lot of switching between the problem and the response alternatives and the consideration of more response alternatives. Thus, he concluded that subjects might follow the response elimination strategy which would involve kind of a backwards approach that aims at the successive elimination of implausible response alternatives.

In a recent study, we discovered that these strategies could be influenced by certain instructions (Loesche, Wiley, & Hasselhorn, 2015). Given the potential importance of detecting hidden rules during inductive reasoning, we manipulated whether test-takers of a matrix reasoning test know the underlying rules by introducing a short teaching session in advance to the test. The results suggested that eye movement behavior under the condition of known rules shifted towards the potentially more efficient strategy of constructive matching. This finding raised a couple of questions that we wanted to address with the experiments presented here. First, we wanted to investigate whether this effect can be replicated by a within-subjects manipulation of rule knowledge. The previous study used a between subjects manipulation and not only found that eye movement patterns differ between experimental groups but also that the correlation with measures of working memory revealed to be more substantial when subjects knew the rules. Testing this hypothesis required a fairly large sample and in the interest of test power, the possibility of a within subjects experiment seemed appealing in order to replicate and further investigate how the correlation of matrix reasoning tests change after providing the relevant rule knowledge.

This leads to our second research question, addressing the issue whether the correlation with other common measures of fluid intelligence (gf) would change when the relevant rule knowledge was provided. Our preferred interpretation of the finding that the correlation with working memory measures increases when the rules are known was that the rule induction process is independent from working memory capacity. We argued that prominent working memory theories generally focus on the management of available information but cannot really explain how entirely new information could be created, which should be a necessary step in solving the items of an inductive reasoning test. For the same reason we predicted to find a decreased correlation with alternative measures of fluid intelligence when the rules are known. That is, if the matrix reasoning test did not share the rule induction process with other measures, then these measures should have less common variance.

The alternative interpretation of said results, however, bears on changes in strategy use that occurred due to knowing the rules. As noted before, the constructive matching strategy is assumed to be more demanding to cognitive resources like working memory capacity, because a mental representation of one or even more potential solutions needs to be created from the multitude of information in the problem. So the pure fact that knowing the rules changes people’s strategies towards this demanding strategy might explain why the correlation with measures of working memory capacity increases under those circumstances while their performance increases at the same time. In that case however, correlations with other measures of gf should be largely unaffected by rule knowledge, or even increase as well.

Eye Movements

Recently, Hayes, Petrov, and Sederberg (2015) pointed out how important the study of eye movements is not only to describe strategy use, but also as an indicator of test score gains in longitudinal cognitive training studies. Eye movements, as an indicator of strategy use, could serve as an objective measure of training effects and would be more clearly interpretable than simple test score gains. Hence, it is important to understand under what circumstances strategies emerge and how they are related to item difficulty or a person’s cognitive ability. Rule knowledge has recently been considered as a potential explanation for secular gains in IQ scores, known as the Flynn Effect (Armstrong & Woodley, 2014; M. C. Fox & Mitchum, 2013), and given the potential link between rule knowledge and strategies (Loesche et al., 2015) this could open up a new methodological approach by studying how secular IQ gains coincide with reasoning strategies.

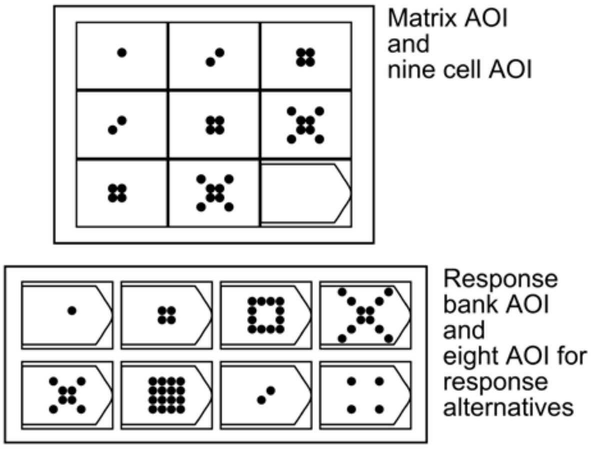


Figure 7. Illustration of a typical Raven item with designated areas of interest (AOI).

Vigneau et al. (2006) presented an extensive analysis of quantifiable eye movement patterns during the solution of the Raven test. They investigated eye movement variables that were mainly defined relative to the response area and the matrix area of the Raven test (Figure 7). That is, these variables reflect the degree to which test-takers seek visual information from the figural problem space or from the response alternatives. It stands to reason that seeking information from the matrix area should coincide with the decomposition of the figural elements and the mapping of these elements to common rules (Carpenter et al., 1990). We will argue that this phase of problem solving behavior serves to build up a mental model (Johnson-Laird, 2005) that can logically explain all the perceived changes and relationships of the items' figural elements and provide at least one potential solution. It should be only after an idea of a potential response comes to mind that a test-taker seeks further information from the response alternatives in order to see if something matches their hypothesis. This solution strategy was labeled constructive matching by Snow (1980) and he characterized this strategy as systematic and effective. A subsequent study suggested that this strategy is more likely to be deployed by test-takers with high reasoning ability (Bethell-Fox et al., 1984). The strategy might not lead to a solution on the first iteration so the test-taker will probably backtrack by seeking more information from the matrix area and this behavior should be visible in eye-movements. Vigneau et al. (2006)

labeled the frequency of macro saccades running between the matrix area and the response area *rate of toggling* and it likely reflects the success rate of the constructive matching strategy, which we assume to be the default strategy by any test taker. Furthermore, Jarosz and Wiley (2012) reported that persons with high working memory capacity tended to exhibit a lower rate of toggling on the Raven test, which could be indicative of more extensive and coherent mental models. Along the same line of research, Hayes, Petrov, and Sederberg (2011) were able to use a new methodological approach to show that the rate of toggling is closely related to overall test performance. They constructed a scanpath successor representation of eye movements on a matrix reasoning test. This is a data driven method that extrapolates regularities in the visual scanpaths by estimating the location of successive fixations under the condition of a current fixation. Two principal components were extracted from this analysis and they were interpreted as an anti-toggle component and a systematicity component. The anti-toggle component corresponds to the already described toggle rate and the systematicity component describes a problem solving behavior of analyzing the matrix area systematically row by row and cell after cell and was closely related to overall test performance, as well.

Vigneau et al. (2006) concluded that eye movements have differing relations depending on whether the analysis is at the individual subject level or at the item level. On the item level, they found that difficulty was “related to solution latency, matrix inspection, response-choice inspection, the number of alternation between interest areas, and latency to first alternation” (p. 270). That is, item difficulty affected almost all the investigated variables. On the individual subject level, they found that better test-takers tend to spend relatively longer time gazing at the matrix area and exhibit less frequent saccades between the matrix area and the response alternatives, which is generally in line with the idea that good test scores coincide with a constructive matching strategy. Given that there might be differing effects depending on the level of analysis, the present study was designed to investigate these effects in multi-level-models. In line with the research reviewed here, we expected to find that item difficulty increase the frequency of response elimination, while ability would facilitate the use of constructive matching. Rule knowledge was expected to work as a facilitator of constructive matching strategy because it provides additional information to incorporate into a mental model.

Response Times

Although not explicitly reported in Loesche et al. (2015), the datasets appended to that paper suggest a strong positive relationship between overall response time and test performance on the Raven test. A reanalysis revealed correlations ranging from .58 in Experiment 4 to .74 in Experiment 2. Thus, one goal of the experiments presented here, was to investigate this relationship further by considering eye movement behavior, item difficulty, and independent measures of ability. The relationship between response time and performance in complex reasoning tasks has already been the subject of investigation in numerous previous studies and they suggest that it is important to differentiate between effects on the item level and effects on the person level.

Goldhammer and Klein Entink (2011), for instance, utilized a joint item response modeling technique that simultaneously accounted for reasoning ability along with reasoning speed and estimated corresponding latent trait parameters. In their sample of 230 students the latent trait estimates for reasoning ability and reasoning speed correlated negatively (r = -.36) meaning that slow test-takers tended to have a higher level of ability. While reasoning ability was also correlated with executive attention, reasoning speed was not, which lead the authors to conclude that there may be different cognitive mechanisms underlying these two latent traits. They assumed that the correlation between reasoning ability and speed can be explained via mental models, by assuming that a good test performance would be grounded in the consideration of alternative solution approaches, especially during the final stage of the mental model building process. This means that able test-takers would question their assumptions before choosing a response, thus taking more time to reach a solution, but having a higher probability of reaching a correct solution. This fits with Doerfler and Hornke (2010) reporting that extraversion lead to lower response latencies on matrix reasoning items, which in turn was detrimental to test performance.

Scherer, Greiff, and Hautamäki (2015) discussed the possible role of motivational factors. They estimated the correlation between task performance and time on task at r = .40 on complex problem solving tasks. Furthermore, their data indicated that motivation (measured with a short questionnaire) may be able to explain some variance in time on task, but not in problem solving ability. This suggests that ability and speed may be somewhat different constructs with differing underlying mechanisms, where motivation is assumed to underlie time.

Goldhammer, Naumann, and Greiff (2015) argued that the degree to which a task is tackled via controlled and automated processes determines the nature of the relationship to response times. If items are easy or subjects are able then the degree of automated processes would be high, thus fast responses could be expected. In a sample of 230 Raven test results, they estimated an overall negative response time effect on the item level in a multi-level model. Item performance on a logit scale was generally worse when response time increased. Furthermore, this fixed effect was negatively correlated with the random intercept across persons, indicating that this effect was even stronger in test-takers of high ability. That is, the gradient in item performance from low to high response times is large in highly able test-takers, and rather flat on the other end of the ability spectrum. This finding seems contrary to previous findings of a positive correlation between time and ability, especially since the sample for this analysis appears to be the same as that of Goldhammer and Klein Entink (2011). But at a second glance it becomes clear that this relationship is dependent on the level of analysis. Goldhammer et al. (2015) essentially found that hard items take longer time (negative correlation between time and score) which actually is just what Goldhammer and Klein Entink (2011) reported for this level of analysis. They found a correlation of r = .63 between an items' difficulty and time intensity (i.e., negative correlation between time and score), but focused the discussion more on the finding of a negative correlation between person level speed and ability. Neubauer (1990) has already noted that the relationship between response time and ability depends on item difficulty and can go in opposite directions on very easy and very hard reasoning items. Thus, we wanted to analyze the covariance patterns between time and performance in multi-level models to clarify the interrelations and the possible implications for the interpretation of eye movements and strategy choices. Response times are potentially relevant in the interpretation of eye movement behavior since most variables are defined in relation to time (e.g., toggles per second or time on matrix). Furthermore, eye movement behavior could shed some light on the effects of ability, difficulty, and rule knowledge on response times, since they could provide information about how exactly people use their time.

Pilot Study

With this pilot study we wanted to test whether it was feasible to realize the rule teaching paradigm in a within-subjects design. There were a couple of concerns connected with it. The biggest issue was that subjects could not unlearn the rules once they have learned about them. Thus, the known-rules condition could only come after the control condition. As a result, whatever effect might occur due to the experimental manipulation could be confounded with test experience, or item characteristics, or other factors such as test motivation and fatigue. To minimize the effects of such potential confounds, the experiment was planned as a 2-by-2-mixed-design with rule knowledge as within subjects factor and item block order counterbalanced across subjects. The Raven APM test (Raven et al., 1998) was split up in two item blocks (A and B) of which one was presented first and was preceded by an introduction video that covered the basic controls and logic of the task and provided four example items. After that, a video explained five rules that govern the elements in the matrix reasoning items followed by the second item block. Thus, there were two item block orders (referred to as AB and BA) varied randomly between subjects and a within-subjects factor of rule knowledge which was introduced between pretest and posttest.

Two hypotheses should be tested with this experiment and the following pattern of results was derived as predictions. First, test performance should significantly improve from rule knowledge. This effect was consistently demonstrated in previous experiments (Loesche et al., 2015) and should confirm that the knowledge of relevant rules has an impact on test performance. Second, eye movements should change from pretest to protest in a way that they reflect more systematic solution strategies like constructive matching.

Task Material

We started out with the same item pool from the Raven test as in Loesche et al. (2015) in order to make the items in the test compatible with the rules being taught in the video. We then split the items up based on odd and even items and made some adjustments by considering item characteristics such as difficulty and the kinds of rules involved (see Table 11). For example, there were a total of seven items in Block A as well as in Block B that contain a constant rule at least once.

Table 11
Raven items in the pilot study, arranged in order of appearance within each block

Item	Block	Any constant rule	Any progress rule	Any one of each rule	Any plus/minus rule
APM 1	A	x		x	
APM 3	A	x	x		
APM 5	A	x	x		
APM 7	A				x
APM 9	A				x
APM 11	A				x
APM 13	A	x		x	
APM 14	A	x	x		
APM 21	A	x		x	
APM 23	A				x
APM 26	A		x	x	
APM 28	A			x	
APM 32	A	x	x		
APM 2	B	x	x		
APM 4	B	x	x		
APM 6	B	x	x		
APM 8	B			x	
APM 10	B	x	x		
APM 12	B				x
APM 16	B				x
APM 17	B	x		x	
APM 22	B				x
APM 24	B	x	x		
APM 27	B			x	
APM 29	B			x	
APM 34	B	x		x	
Sum Block A:		7	5	5	4
Sum Block B:		7	5	5	3

Eye Movement Measures

There are multiple ways to summarize eye movement data. Following Vigneau et al. (2006) we defined variables with respect to two interest areas: the matrix and the response bank (Figure 7). The toggle rate (TR) is based on the number of saccades running either way between the matrix area and the response area (number of toggles) divided by total response time in seconds (i.e., toggles per second). The relative time on matrix (RTM) is the summed duration of all fixations within the matrix area (time on matrix) divided by total response time. The relative first response fixation (RFRF) is the timestamp of the first fixation within the response area (time before first response fixation) divided by total response time. Following Bethell-Fox et al. (1984) we also investigated the total number of responses visited (RV) and counted all response alternatives that were fixated at least once during the response time.

Procedure

A total of 50 undergraduate subjects (mean age = 24 years) were randomly assigned to either the AB (*n* = 28) or BA (*n* = 22) item block order condition. Two additional participants were excluded due to technical issues or low quality of eye movement data. Participants were tested in single sessions lasting between 30 and 45 minutes and received money or course credit in exchange. An EyeLink 1000 device recorded eye movements at a sample rate of 500 Hz. Where necessary, we performed post-hoc adjustments to account for obvious miss-calibrations by aligning whole fixation patterns to the stimuli. After receiving information about the purpose of the experiment the eye tracker was calibrated and participants started working on either block A or block B, depending on their random assignment. The first block was preceded by four practice items and an introduction video of 2:28 minute length. Noteworthy, this video was revised from the one described in Loesche et al. (2015) in order to iron out some discrepancies between testing conditions. The example items shown in this video were the same as in the rule teaching video and correct responses to all example items were given. However, it was not explained why the responses were correct and neither was any rule mentioned. This was to ensure that differences between instructions were only attributable to whether rules were taught or not. After watching this video, participants worked on the first 13 items with no time restrictions while eye movements were recorded. Each item trial started with the presentation of a fixation cross, designed after recommendations from Thaler, Schütz, Goodale, and

Gegenfurtner (2013), to ensure the fixation was at the center of the matrix area. Responses were indicated with the mouse and there was no time limit and no feedback on the response. After the first block, a rule teaching video (4:52 minute) was presented, which was basically the same as in Loesche et al. (2015), except that some part was trimmed from the beginning since subjects were already familiar with the basics of the task. The video showed the same four example items as in the pretest instruction and each item exemplified one or two of the five rules labeled as *Plus, Minus, Constant, Progress,* and *One of Each*. Then participants worked on 13 items from the second block and finally answered a free recall question about the rules they could remember from the rule teaching video.

Prior to working on the Raven test, we asked participants to complete a backward digit span task, which consisted of 14 items. On each consecutive item there was a series of single digits with increasing lengths from 2 to 8. Each digit was presented for 500 ms and afterwards participants were asked to recall the digits in reverse order within a 15 second response time window.

Results

We briefly report the basic results of the pilot study and go into more detailed analyses with the data from the main experiment. The online supplementary material contains complete datasets and scripts for R (R Core Team, 2015), not only for the pilot study but also for the main experiment, so as to allow the interested reader to retrace and extend our analyses (see Appendix A).

Table 12
Dependent variable means in the pilot study across two factors

Rule knowledge	Block order	Raven score	Raven log(time)	RTM	RFRF	TR	RV
Pre	AB	0.574 (0.176)	3.353 (0.371)	0.667 (0.054)	0.436 (0.115)	0.407 (0.131)	5.648 (1.278)
Post	AB	0.709 (0.144)	3.476 (0.237)	0.692 (0.083)	0.466 (0.140)	0.334 (0.132)	5.349 (1.157)
Pre	BA	0.731 (0.108)	3.331 (0.365)	0.686 (0.052)	0.423 (0.119)	0.356 (0.117)	5.759 (0.810)
Post	BA	0.699 (0.150)	3.609 (0.240)	0.713 (0.050)	0.517 (0.106)	0.292 (0.094)	5.601 (1.071)

Note. Values in parentheses are *SD*. RTM = relative time on matrix, RFRF = relative first response fixation, TR = toggle rate, RV = responses visited, MTDI = matrix time distribution index.

As a confirmation that the manipulation worked we expected to find that subjects would perform better after the rules were taught, which was generally confirmed by significant within subjects main effect, *F*(1, 48) = 13.18, *p* < .01. However, this effect was qualified by a significant interaction with block order, *F*(1, 48) = 23.65, *p* < .01. The pattern was similar for log-transformed response time for which the main effect of rule knowledge, *F*(1, 48) = 32.65, *p* = <.01, interacted with block order, *F*(1, 48) = 5.22, *p* = .03. As can be seen in Table 12, the two block order conditions already differed significantly on the pretest, *F*(1, 48) = 13.43, *p* < .01. Performance on the backward digit span was comparable between conditions, *F*(1, 47) = 0.68, *p* = .41. This implies that item compositions had different difficulty levels.

Furthermore, there were significant main effects of rule knowledge on the relative time on matrix, *F*(1, 48) = 10.36, *p* < .01, on toggle rate, *F*(1, 48) = 43.53, *p* < .01, and on the relative first response fixation, *F*(1, 48) = 16.41, *p* < .01. These effects were consistent with the hypothesis that strategy shifted towards constructive matching. The effect on the relative first response fixation was also qualified by a significant interaction effect, *F*(1, 48) = 4.83, *p* = .03. The amount of responses visited did not change significantly due to rule knowledge, *F*(1, 48), *p* = .11.

Main Experiment

The results of the pilot study suggest that the experimental manipulation had an effect on most of the investigated eye movement variables and partly on test performance. However, the results may have been biased by different levels of difficulty between item blocks. In this subsequent experiment we tried to address this issue by revising the composition of item blocks. Additionally, we wanted to investigate the relationship to other measures of fluid intelligence and whether the relationship would change due to the experimental manipulation. The correlation with working memory measures was rising with rule knowledge in Loesche et al. (2015). This effect could have been due to an increased relative task demand on working memory because a test like the Raven would essentially turn into a measure of working memory capacity when there is no need to come up with possible rules. Hence, we expected to find an increased correlation with the backward digit span after the rules were known.

But the main purpose of the current experiment was to test the reverse logic of this hypothesis. This was based on the assumption that many other measures of fluid intelligence have in common with the Raven test, the requirement to detect underlying rules of the task material. So if the impact of the ability to come up with rules was eliminated from the Raven test, then the test should have less in common with other measures of the like. Hence, we predicted to find a reduced correlation with measures of fluid intelligence when the rules were known on the posttest. An alternative explanation for the finding of an increased correlation with working memory was connected with the finding that strategies shifted towards constructive matching. This strategy is theoretically more demanding to information processing entities like the working memory system, so if test-takers rely more frequently on this strategy they are more likely to reach the limits of working memory. This could as well explain why the correlation with measures of working memory increases under the condition of known rules, but the correlation with measures of fluid intelligence should remain unaffected.

Finally, we wanted to explore the effects of rule-knowledge on response time. As pointed out in the introduction, previous research suggests a strong positive relationship on the person level. Hence, we expected to find the same positive correlation between response time and test performance. Given the assumption that test performance and test time are positively correlated, it seems plausible to assume that test time under the condition of known rules increases along with test performance. Additionally, using a constructive matching strategy might theoretically take more time since it is the more laborious way. On the other hand, it might also be the more efficient way, thus reducing response time. A re-analysis on the dataset from Loesche et al. (2015) revealed that the effects of rule knowledge on response times were weak or nonexistent, with a tendency to increase response time. In the pilot study we found a main effect of rule-knowledge on response time but the effect was limited by an interaction effect. Thus, it was not clear what to expect from the effect of rule knowledge on response time. However, response times are also of potential relevance for the interpretation of eye-movement behavior. Most eye-movement indicators are defined relative to time (i.e., relative time on matrix or toggles per second) but can also be defined as absolute values (i.e., absolute time on matrix or number of toggles). It is important to understand how response times behave in order to understand the meaning of these variables. For example, an increase in the number of toggles per seconds can be due to reduced response time with a constant number of toggles or it can be due to an increased number of toggles in the same amount of time. Furthermore, eye-movements can potentially provide information about the use of time. Snow (1980) hypothesized that response elimination is a fallback strategy that participants would revert to when constructive matching fails. In that case we expect to see an increase of corresponding eye-movement indicators along with an increase in response time.

Task Material

For the main experiment, we revised the item composition of the two blocks to address the issues occurring in the pilot study. We excluded items that were apparently too easy (solution rate of 100%) and item 29 since it seemed very hard and, on a second thought, was not completely compatible with the rule set in the intervention. We finally added some items from the Raven Standard Progressive Matrices (Styles, Raven, & Raven, 1998) to fill up the void. Where possible, item pairs were identified based on difficulty and the kind and number of rules involved (see Table 13). Two items were considered parallel if they required roughly the same number and kind of rules. For example, the first item in APM (APM 1) requires three *one of each* rules. Item number 13 of the same test (APM 13) was considered parallel since it requires one *constant* rule and two *one of each* rules. It was not possible to identify a parallel item for all items in the test, thus the remaining items were allocated to one of the two blocks based on the solution rate obtained in the pilot study.

General Ability Measures

Participants completed a set of three tasks that can be considered as measures of general fluid intelligence. We used a figural analogies test with 15 items, loosely adapted from an unpublished test described in Chuderski, Taraday, Nęcka, and Smoleń (2012). Each item consisted of a figural analogy of the form A:B::C:D where D is a missing figure that needs to be chosen from five response alternatives. Due to a graphical error one of the items in this test (number six) was ambiguous and therefore excluded from all analyses. We also used a letter sets test with 15 items based on the same test in the *Kit of Factor-Referenced Cognitive Tests* (Ekstrom, French, Harman, & Diran, 1976). Each item in this test consisted of five groups with four letters in each group. The letters in the groups were arranged based on some hidden rule whereas one group violated this rule. The subjects were asked to identify the one letter set that did not fit the rule. As a third measure of gf we utilized a number series test, similar to the one described in Thurstone (1938). Each of the 15 items in this test displays a series of seven numbers that follow some hidden rule. The participants were asked to complete the series by entering one number that would correctly complete the series. This task did not require a multiple choice response, but a free response that required more than one button click, which might have affected response time measures (see Results section).

Table 13
Items selected for the main experiment, arranged in order of appearance within each block

Original item number	Pilot study pretest solution rate	Selected for Block	Roughly parallel to
APM 3	0.852	A	APM 2
APM 5	0.889	A	APM 4
APM 9	0.926	A	APM 7
APM 1	0.741	A	APM 13
SPM D2	--	A	APM 12
SPM E1	--	A	APM 14
APM 16	0.909	A	--
APM 22	0.409	A	APM 23
APM 21	0.370	A	SPM E2
SPM E7	--	A	APM 24
APM 26	0.296	A	--
APM 34	0.727	A	APM 28
SPM E4	--	A	--
APM 32	0.222	A	SPM E9
APM 2	0.864	B	APM 3
APM 4	0.909	B	APM 5
APM 7	0.741	B	APM 9
APM 13	0.481	B	APM 1
APM 12	0.864	B	SPM D2
APM 14	0.778	B	SPM E1
APM 17	0.864	B	--
APM 23	0.222	B	APM 22
SPM E2	--	B	APM 21
APM 24	0.227	B	SPM E7
APM 27	0.591	B	--
APM 28	0.185	B	APM 34
SPM D8	--	B	--
SPM E9	--	B	APM 32
APM 6	1.000	excluded	
APM 8	1.000	excluded	
APM 10	1.000	excluded	
APM 11	0.889	excluded	
APM 29	0.136	excluded	

Note. APM = item from Raven’s Advanced Progressive Matrices. SPM = item from Raven’s Standard Progressive Matrices.

Eye Movement Measures

We analyzed the same eye movement indicators as in the pilot study and also broke them down into their numerators and denominators for a more detailed analysis. In this Experiment, we also investigated the matrix time distribution index (MTDI) as defined by Vigneau et al. (2006). This index reflects the relative frequency of fixations at the last column and the last row of the matrix area and it was calculated as time of fixations on four top-left cells (1, 2, 4, 5), minus the time of fixations on the other five cells (3, 6, 7, 8, 9), divided by the total time on the matrix area (Figure 7). Finally, we suspected that our experimental manipulation might have had an effect on the direction of the matrix inspection. Both the initial instruction and the rule teaching instruction emphasized to analyze the matrix area horizontally, but the rule instructions were all explained row-wise so they might have given test-takers additional conviction that this was the right approach. Thus, we counted all horizontal saccades (HS) within the matrix area as a proportion of the total number of saccades in the same area.

Procedure

A total of 109 undergraduate students (mean age = 24 years) participated in this study in exchange for money or course credit. Four participants were excluded from analyses pertaining to the Raven test due to technical problems during testing, the remaining participants were randomly assigned to either the AB item block order (*n* = 56) or the BA item block order (*n* = 49). One additional participant was excluded from analyses pertaining to eye movements because the recorded data was distorted and too noisy. Participants started the testing session with three measures of gf in randomized order, followed by the same backward digit span task as in the Pilot study. After that, the experiment followed the same procedure as in the pilot study. Testing sessions usually lasted less than one and a half hours.

Results

Analyses of Variance. First, we wanted to confirm that the experimental manipulation resulted in an increased test performance and that the two item blocks were indeed parallel and thus comparable. A two-way ANOVA with item block order as between subjects factor and rule knowledge as within subjects factor, revealed a significant main effect of rule knowledge, *F*(1, 103) = 17.97, p < .01, which was again qualified by a significant interaction with item block order, *F*(1, 103) = 8.75, p < .01. In contrast to the pilot study, pretest performances did not differ significantly on the two item blocks, *F*(1, 103) = 1.42, p = .24. This suggests that item blocks were much more comparable than in the pilot study but still not completely parallel. There was no main effect of rule knowledge on log-transformed response times, *F*(1, 103) = 2.70, p = .10, there was however a significant cross interaction effect, *F*(1, 103) = 89.99, p < .01, indicating that response times were independent of our experimental intervention and were rather tied to the specific item compositions (Table 14). A comparison of the gf estimates for each block order group, suggested that they were comparable, *F*(1, 103) = 0.23, p = .63.

Table 14
Dependent variable means in the main experiment across two factors

Rule knowledge	Block order	Raven score	Raven log(time)	RTM	RFRF	TR	RV	HS
Pre	AB	0.669 (0.150)	3.271 (0.363)	0.808 (0.04)	0.481 (0.116)	0.328 (0.071)	5.812 (0.676)	0.758 (0.103)
Post	AB	0.690 (0.180)	3.484 (0.353)	0.815 (0.04)	0.478 (0.095)	0.304 (0.072)	5.975 (0.708)	0.807 (0.051)
Pre	BA	0.634 (0.177)	3.422 (0.353)	0.797 (0.045)	0.432 (0.129)	0.338 (0.109)	6.288 (0.604)	0.761 (0.097)
Post	BA	0.730 (0.190)	3.251 (0.325)	0.827 (0.042)	0.520 (0.137)	0.297 (0.091)	5.330 (0.847)	0.805 (0.053)

Note. Values in parentheses are *SD*. RTM = relative time on matrix, RFRF = relative first response fixation, TR = toggle rate, RV = responses visited, HS = horizontal saccades.

All eye movement variables were significantly affected by the within subjects factor rule knowledge and all effects pointed in the predicted directions. Relative time on matrix was larger with rule knowledge, *F*(1, 102) = 27.16, *p* < .01. The first response fixation occurred later with rule knowledge, *F*(1, 102) = 17.42 , *p* < .01. The toggle rate was lower with rule knowledge, *F*(1, 102) = 22.93, *p* < .01. The number of responses visited was lower with rule knowledge, *F*(1, 102) = 37.86, *p* < .01. Parallel to test performance, three effects were qualified by significant interaction effects with item block order. This was the case with the relative time on matrix, *F*(1, 102) = 11.61, *p* < .01, with the first response fixation, *F*(1, 102) = 23.59, *p* < .01, and with the number of responses visited, *F*(1, 102) = 88.92, *p* < .01. There was no interaction effect on toggle rate, *F*(1, 102) = 1.62, *p* = .20.

Correlations. Table 15 shows the correlations between measures of gf and the two blocks of the Raven test collapsed across the two order conditions. We calculated paired sample significance tests (Revelle, 2015) on the differences of correlations between testing conditions. The results suggest that Raven scores did not correlate differently with any of the tested measures after the rules were given to participants (all p > .38).

Table 15
Correlations of Raven test scores with other ability measures.

	Pre Raven	Post Raven	Figural Analogies	Letter Sets	Number Series
Post Raven	.691**				
Figural Analogies	.536**	.593**			
Letter Sets	.456**	.510**	.655**		
Number Series	.472**	.443**	.602**	.523**	
Backward Digit Span	.308**	.330**	.164	.198*	.301**

Note. *N* = 105. ** *p* < .01. * *p* < .05.

Table 16
Population estimates of the correlation between person speed and ability.

	Speed						
	Raven Pretest Block A	Raven Posttest Block B	Raven Pretest Bock B	Raven Posttest Block A	Figural Analogies	Letter Sets	Number Series
Ability	-.482	.022	-.655	-.219	-.667	-.732	-.074

Conjoint Item Response and Time Models. Response times were further investigated by utilizing a conjoint item response and time modeling approach (CIRT) that simultaneously estimates item and person parameters for response time and test performance (J.-P. Fox, Entink, & van der Linden, 2007). Note that the CIRT models operate with a speed variable instead of a response time variable, meaning that the model algorithms multiply all response times with negative one. This analysis revealed that population estimates for speed and ability were mostly negatively correlated (see Table 16). That is, subjects who were overall slower on the test were more likely to achieve a higher score, and vice versa. There were two exceptions. First, the correlation between ability and speed for the number series task was almost zero and this might be due to the response format in this task. Unlike all the other tasks the number series response required more than just a single click and was not multiple-choice. Thus, it seems plausible to assume that the requirement to construct the response from nine digits, added additional variance to the response time data. Responses on the other two gf tasks required only one click, thus times should closely reflect cognitive processes affected by the task itself. Second, the correlation between speed and ability on the Raven dropped considerably from pretest to posttest to almost zero in group AB and to -.219 in group BA. Noteworthy this drop was about the same magnitude in both experimental conditions and may have occurred due to the rule knowledge manipulation.

Table 17 displays the complete correlation table of ability and speed for the Raven and the fluid intelligence measures. These are not the population correlation estimates from the CIRT models reported earlier. Instead, all person parameter estimates were attributed to the corresponding subjects and sample correlations were computed based on these estimates. This allows the examination of correlations between ability and speed estimates across different tasks. Considering the ability-speed correlations, it strikes that they were negative within all tasks, meaning that with increasing response times, task performance became better, and vice versa. Correlations of ability and speed within the same fluid intelligence tasks were r = -.81 for figural analogies and r = -.84 for letter sets, but only r = -.15 for number series. As mentioned above, this was likely due to the response modality in this task. In line with this interpretation is also the fact, that response times on figural analogies and letter sets explain a good amount of variance in number series task performance.

On the Raven pretest correlations of ability and speed were r = -.57 and r = -.74 in the two conditions. However these numbers change drastically on the posttest, where correlations drop to .02 and -.25. That is, the correlation drops by about 50 points in both item order conditions from pretest to posttest, so this effect occurs regardless of the subsample and the items (both $\Delta p < .01$). That is, variance in time did not explain as much variance in performance after the rules were known and this finding corresponds to the population correlation estimates from the CIRT model (see above).

Finally, Table 17 reveals that time and performances were correlated across tasks, suggesting that there was not only an individual disposition for ability in this sample, but also for speed. For example, speed on the letter sets test was correlated with Raven test performance between r = -.19 and r = -.59, depending on the experimental condition, and there were similar cross-task relationships between ability and speed with other task combinations.

Table 17
Sample correlations of person level estimates for speed and ability from the CIRT models.

	Ability				Speed (-time)				Ability			Speed (-time)		
	AB		BA		AB		BA		Gf					
	1) Pre	2) Post	3) Pre	4) Post	5) Pre	6) Post	7) Pre	8) Post	9) FA	10) LS	11) NS	12) FA	13) LS	14) NS
1)	1													
2)	.669	1												
3)			1											
4)			.810	1										
5)	-.570	-.251			1									
6)	-.271	.022			.790	1								
7)			-.739	-.474			1							
8)			-.449	-.250			.796	1						
9)	.631	.585	.663	.615	-.510	-.259	-.624	-.437	1					
10)	.490	.332	.621	.599	-.558	-.359	-.621	-.501	.703	1				
11)	.506	.445	.460	.421	-.411	-.091	-.465	-.274	.619	.544	1			
12)	-.548	-.339	-.554	-.448	.653	.519	.718	.728	-.758	-.682	-.462	1		
13)	-.420	-.186	-.594	-.555	.664	.564	.666	.563	-.629	-.830	-.431	.780	1	
14)	-.302	-.096	-.438	-.355	.558	.532	.611	.660	-.384	-.454	-.135	.705	.657	1

Note. FA = figural analogies, LS = letter sets, NS = number series, AB and BA refer to block-order conditions. *N* = 105, n_{AB} = 56, n_{BA} = 49.

Next, we report the CIRT model estimates for the item parameter estimates. On the item level, the correlation between population estimates of time intensity and difficulty were generally positive within the tasks in the experiment, meaning that hard items required more time while easy items were solved quickly (see Table 18). This effect may seem contradictory to the person level effect because on the item level it means, when a person has high ability, the test items should be relatively easy for them so they should have responded quickly (implying a positive ability-speed relation). Yet, the actual finding on the person level was that high ability was associated with slow response times (negative ability-speed relation). This finding hints towards distinct relationships between time and performance depending on the level of observation, which we will discuss more thoroughly in the next section. It is also worth mentioning that, unlike the speed-ability correlation, the correlation between difficulty and time intensity is relatively stable between pretest and posttest conditions.

Table 18
Population estimates of the correlation between item parameters for time intensity and difficulty.

	Item time intensity						
	Raven Pretest Block A	Raven Posttest Block B	Raven Pretest Bock B	Raven Posttest Block A	Figural Analogies	Letter Sets	Number Series
Item difficulty	.443	.391	.382	.381	.162	.326	.144

Multi-level Analyses of Performance. To further explore the effect of the experimental manipulation and some further variables on item performance, a series of multi-level models were evaluated. All multi-level analyses were computed with version 1.1.10 of the lme4 R-package (Bates, Mächler, Bolker, & Walker, 2015). We provide p-values based on the Satterthwaite approximation from the lmerTest R-package, version 2.0.29 (Kuznetsova, Brockhoff, & Christensen, 2015) but would like to point out that providing p-values for nested effects is currently subject of debate because of the difficulty of attributing degrees of freedom to the model levels (Bates et al., 2015; Bolker, 2015). Hence, p-values should be interpreted with caution but do provide some guiding perspective on the likelihood of the data presented here.

Table 19
Series of multi-level models predicting the dichotomous dependent variable item performance.

	Predictor variables									
Variance components, and model fit	Rule knowledge	Person ability	Item time	Item difficulty	RFRF	RTM	TR	HS	RV	MTDI
Intercept	1.123** (0.358)	1.325*** (0.344)	1.373*** (0.307)	1.321*** (0.131)	1.336*** (0.340)	1.310*** (0.353)	1.310*** (0.354)	1.319*** (0.349)	1.351*** (0.330)	1.330*** (0.352)
Fixed effect	0.433*** (0.101) [0.236, 0.645]	0.738*** (0.091) [0.564, 0.912]	-0.603*** (0.114) [-0.820, -0.404]	-1.729*** (0.093) [-1.929, -1.535]	1.220*** (0.241) [0.723, 1.739]	1.523* (0.702) [0.185, 2.804]	-0.972* (0.392) [-1.739, -0.214]	1.568** (0.579) [0.484, 2.556]	-0.194*** (0.032) [-0.261, -0.130]	0.122 (0.299) [-0.414, 0.659]
Intercept variance across subjects	1.163	0.568	1.215	1.128	1.138	1.100	1.095	1.121	1.157	1.128
Intercept variance across items	3.090	3.042	2.184	0.067	2.800	3.062	3.080	2.984	2.597	3.033
Number of estimated parameters	4	4	4	4	4	4	4	4	4	4
-2 log-likelihood	2693.82	2659.39	2684.75	2630.23	2651.70	2673.05	2670.09	2670.54	2625.03	2664.81

Note. Each column represents a unique model with a different predictor. Values in parentheses are standard errors. Values in brackets are 95% bootstrap CI. *K* = 28 items were crossed with *N* = 105 subjects. P-values are based on the Satterthwaite approximation of degrees of freedom (see text). RFRF = relative first response fixations, RTM = relative time on matrix, TR = toggle rate, HS = horizontal saccades, RV = responses visited, MTDI = matrix time distribution index. ****p* < .001. ***p* < .01. **p* < .05.

We investigated how various variables predict item performance by estimating a series of logistic multi-level models, each with a different predictor. Three measures of general fluid ability (gf) were averaged and standardizing to serve as an estimate of ability. All eye movement variables were centered to the individual mean. A random effect was specified for the subject level factor, and a further random effect was specified for the item content factor. The item content factor was coded based on the content of the item, irrespective of whether the item was presented at pretest or posttest in a given subsample. Both factors were completely crossed, i.e. there was one observation for any combination of the two factors values. We fitted a series of models, each with a different predictor and compared the model fit to a null model (see Table 19).

The analysis revealed that there was a significant positive effect of rule knowledge on item performance (*β* = 0.433), indicating that the experimental manipulation worked as hypothesized. Person ability and item difficulty were also significant predictors of item performance, but more importantly, there were differences among eye movement variables in their predictive power. Judging by the p-values of the effects, it appears that RFRF (β = 1.220, *p* < .001) and RV (β = -0.194, *p* < .001) are the best single predictors. However, none of the estimated 95% CI overlap with zero, indicating that all eye movement variables were in fact correlated with item performance. The one exception was MTDI, which we will discuss in more detail in the result-section on eye movements.

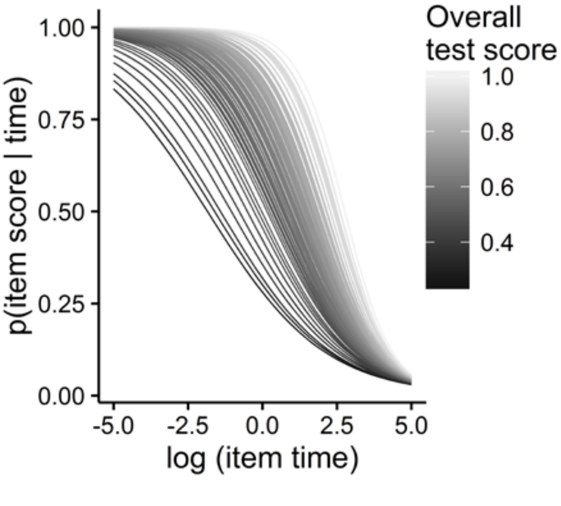


Figure 8. Estimated item score as a logistic function of response time for different levels of subject test performance.

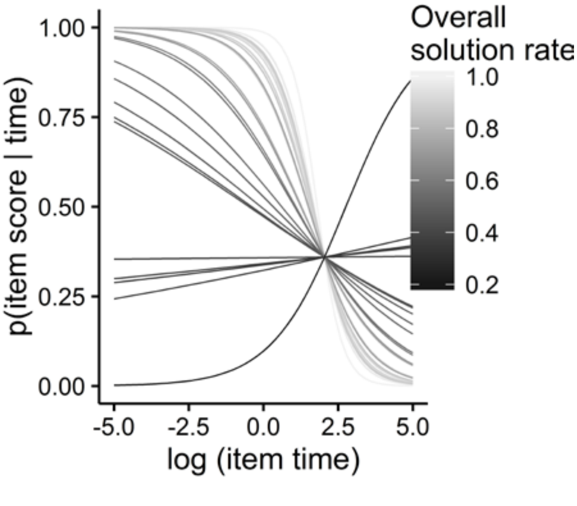


Figure 9. Estimated item score as a logistic function of response time for different levels of item solution rates.

Item response time had a negative effect (β = -.603) on item performance, meaning that items with long response times tended to have a lower probability of a correct response which replicated a finding reported by Goldhammer et al. (2015). We followed their example and specified two additional random effects for the slopes of the effect of time on response across the two grouping factors for subjects and items. This model resulted in a significantly better model fit than the fixed slope model (-2-log-likelihood = 2655, $\Delta\chi^2$ = 29.82, Δdf= 4, $\Delta p < .001$). The random intercept model provides an estimate for the correlation between intercept and slope across subjects. With -1.00 this estimate suggests a perfectly negative relationship, meaning that the negative effect of time on accuracy becomes even more negative with higher intercept levels. Higher intercept levels suggest overall better test performance of a subject, thus it appears that subjects with high ability express a stronger gradient of item score between occasions where they take a lot of time and occasions where they take less time. This is depicted graphically in Figure 8 which shows how subjects with high ability (i.e., high test score) had a steeper estimated decline in the probability to solve an item with increasing time. The same perfectly negative linear relationship was estimated across items, meaning that the negative effect of time on accuracy becomes even more negative with higher solution rates (on easier items). The graphic also shows that the relationship was estimated negative for all subjects, only somewhat weaker for low ability subjects. Figure 9 shows how the estimated probability to solve easy items (i.e., high solution rates) drops significantly with longer response times, but this relationship weakens and even turns around for harder items (i.e., low solutions rates). That is, the probability to solve very hard items is higher with long response times but for most items the probability is lower with long response times. It should be noted however, that estimates of exactly negative one are likely to be a sign that there was not enough information in the dataset for an accurate estimation. That is, there might not be enough levels of one of the random effects or not enough observations.

Table 20
Series of multi-level models predicting item response time

Predictors, variance components, and model fit	Null model	Fixed slope model	Random slope model	Cross-level interaction model 1	Cross-level interaction model 2
Intercept	3.358*** (0.108)	3.358*** (0.063)	3.358*** (0.070)	3.358*** (0.063)	3.358*** (0.063)
Person ability		0.119*** (0.030)	0.109** (0.038)	0.119*** (0.033)	0.119*** (0.033)
Item difficulty		0.473*** (0.057)	0.307*** (0.054)	0.473*** (0.058)	0.473*** (0.057)
Rule knowledge		-0.010 (0.009)	-0.011 (0.008)	-0.011 (0.008)	-0.011 (0.008)
Ability × difficulty				0.124*** (0.019)	0.124*** (0.019)
Ability × rule knowledge					0.030*** (0.008)
Residual variance	0.213	0.213	0.182	0.183	0.182
Intercept variance across subjects	0.102	0.088	0.089	0.089	0.089
Intercept variance across items	0.300	0.084	0.111	0.085	0.085
Slope variance of difficulty across subjects			0.012	0.011	0.011
Slope variance of ability across items			0.016	0.005	0.005
Intercept-slope correlation across subjects			0.185	0.184	0.183
Intercept-slope correlation across items			0.654	0.517	0.528
Number of estimated parameters	4	7	11	12	13
-2 log-likelihood	4220.51	4169.79	3887.73	3857.69	3844.05
compared to previous model		$\chi^2(3) =$ 50.72***	$\chi^2(4) =$ 282.06***	$\chi^2(1) =$ 30.04***	$\chi^2(1) =$ 13.64***
ICC	0.448				

Note. Values in parentheses are standard errors. *K* = 28 items are crossed with *N* = 105 subjects. P-values are based on the Satterthwaite approximation of degrees of freedom (see text). ***p < .001. **p < .01. *p < .05.

Multi-level Analyses of Time. In order to better understand the slope-intercept correlation, it was beneficial to interchange the predictor and the dependent variable in the previously presented models by predicting item response time from accuracy. This could overcome limitations of predicting a dichotomous variable and the problems to estimate random effect covariance in the previous models. We were interested in how item difficulty (level one) and person ability (level two) affected item response time. Three measures of general fluid ability (gf) were averaged and standardizing to serve as an estimate of ability. As a measure of item difficulty we used the estimates from the CIRT model and standardized the values to make them comparable to the level two measure of general ability. We used only pretest estimates and expanded these to posttest items because they were unaffected by our experimental manipulation. This was warranted by the fact that both subsamples did not differ in terms of gf, *F*(1, 103) = 0.23, p = .63. A random effect was specified for the subject level factor and a further random effect was specified for the item content factor. Both factors were completely crossed, i.e. there was one observation for any combination of the two factor's values.

The random intercept and fixed slope model suggests that person ability and item difficulty explain a significant amount of variance in item response time (see Table 20). Item difficulty had about four times as strong of an effect compared to ability. That is, for each standard deviation increase in item difficulty the response time for that item increases by about 0.473 log(seconds), while a standard deviation increase in ability increases response time by 0.119 log(seconds). The rule teaching intervention did not seem to have an effect on response time, which confirms the ANOVA result reported above.

The random slope model fits the data significantly better than the fixed slope model as indicated by a likelihood ratio test (see Table 20). This indicates that slopes have a variance that is significantly different from zero and allows for the test of interactions between fixed effects (Aguinis, Gottfredson, & Culpepper, 2013). The cross-level interaction model estimates the interaction effect between person ability and item difficulty at $\beta = 0.133$, meaning that subjects with high ability took even more time on hard items than subjects with low ability. This is depicted graphically in Figure 10 illustrating that, on very easy items, subjects with low ability tended to take slightly longer than their counterparts. The opposite was true at the other end of the item difficulty spectrum, where response times of low ability subjects fall behind those of high ability subjects. It appears that high ability subjects were slightly faster than low ability subjects on very easy items, but adapted to the difficulty level of the items by taking more time on hard items. On the other hand, low ability subjects were slightly slower on easy items and jumped to hasty conclusions on hard items, which were more likely to be wrong. Although not reported here, we found similar results for the figural analogies task and for the letter sets task (see supplementary material).

There was also a significant negative interaction between the effects of ability and the intervention (β = -0.059), suggesting that the relationship between ability and response time is significantly weaker with rule knowledge (see Figure 11). This confirms the results from the CIRT models and a visual inspection of the interaction effect suggest that low ability subjects increased their response times with rule knowledge, while high ability subjects showed barely any difference between pretest and posttest measures.

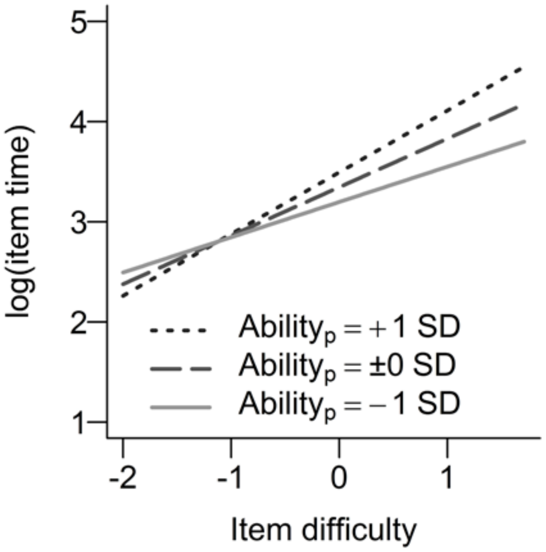


Figure 10. Estimated interaction effect of item difficulty and person ability on item response time. The x-axis covers the range of values in the sample.

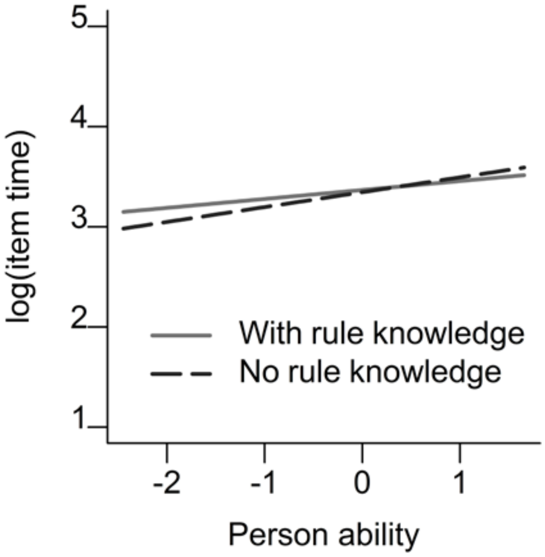


Figure 11. Estimated interaction effect of person ability and rule knowledge on item response time. The x-axis covers the range of values in the sample.

Table 21
Series of multi-level models predicting various eye movement indicators.

Predictors, variance components, and model fit	TR	NTOG	RTM	TM	TA	RFRF	BFRF	AFRF
Intercept	0.332*** (0.013)	11.418*** (0.666)	0.803*** (0.010)	31.268*** (1.989)	7.052*** (0.491)	0.457*** (0.020)	14.302*** (0.947)	24.038*** (1.795)
Person ability	-0.022** (0.007)	0.657 (0.420)	0.013*** (0.003)	5.717*** (1.274)	0.442 (0.251)	0.003 (0.011)	2.244*** (0.618)	3.841** (1.188)
Item difficulty	-0.018 (0.011)	3.993*** (0.574)	0.008 (0.009)	14.406*** (1.821)	2.640*** (0.461)	-0.066*** (0.018)	5.017*** (0.833)	12.051*** (1.618)
Rule knowledge	-0.032*** (0.005)	-0.793** (0.268)	0.018*** (0.003)	1.286 (0.668)	-0.396* (0.173)	0.042*** (0.008)	1.591*** (0.395)	-0.746 (0.707)
Ability × difficulty		0.904** (0.266)		5.646*** (1.005)	0.673*** (0.188)		1.871*** (0.447)	4.398*** (0.931)
Ability × rule knowledge								
Residual variance	0.135	51.503	0.073	17.914	4.633	0.215	10.577	18.927
Intercept variance across subjects	0.005	14.062	0.001	96.066	4.384	0.010	26.584	89.235
Intercept variance across items	0.003	7.632	0.002	78.668	5.141	0.008	15.761	59.238
Slope variance of difficulty across subjects		2.962		29.878	1.465		7.209	30.645
Slope variance of ability across items		0.631		16.366	0.365		2.421	11.903
Intercept-slope correlation across subjects		0.969		0.989	1		1	1
Intercept-slope correlation across items		0.364		0.872	0.314		0.816	0.801
Number of estimated parameters	7	12	7	12	12	7	12	12
-2 log-likelihood	-3050.27	19943.34	-6613.48	25314.12	17441.82	-395.96	22112.11	25488.36
ICC	0.323	0.398	0.386	0.501	0.404	0.324	0.357	0.412

Note. Table continues on next page. Each column represents a unique model with a different dependent variable. Values in parentheses are standard errors. K = 28 items were crossed with N = 105 subjects. P-values are based on the Satterthwaite approximation of degrees of freedom (see text). TR = toggle rate, NTOG = number of toggles, RTM = relative time on matrix, TM = time on matrix, TA = time on alternatives, RFRF = relative first response fixations, BFRF = time before first response fixation, AFRF = time after first response fixation, RV = responses visited, HS = horizontal saccades, MTDI = matrix time distribution index, MTDSD = matrix time distribution standard deviation. ***p < .001. **p < .01. *p < .05.

Table 21 (continued)
Series of multi-level models predicting various eye movement indicators.

Predictors, variance components, and model fit	RV	HS	MTDI	Abs(MTDI)	MTDSD
Intercept	6.063*** (0.211)	0.759*** (0.008)	0.030 (0.028)	0.201*** (0.008)	7.067*** (0.061)
Person ability	0.072 (0.073)	0.003 (0.007)	0.031** (0.011)	-0.004 (0.005)	0.103** (0.035)
Item difficulty	0.783*** (0.206)	-0.010 (0.005)	-0.014 (0.026)	-0.027*** (0.007)	0.399*** (0.055)
Rule knowledge	-0.399*** (0.061)	0.047*** (0.003)	-0.023*** (0.003)	0.005* (0.003)	-0.028** (0.009)
Ability × difficulty	0.144* (0.053)				0.117*** (0.021)
Ability × rule knowledge					0.031** (0.010)
Residual variance	1.641	0.090	0.032	0.020	0.254
Intercept variance across subjects	0.308	0.004	0.012	0.002	0.096
Intercept variance across items	1.107	0.001	0.018	0.001	0.076
Slope variance of difficulty across subjects	0.034				0.011
Slope variance of ability across items	0.041				0.006
Intercept-slope correlation across subjects					-0.009
Intercept-slope correlation across items					0.312
Number of estimated parameters	12	7	7	7	13
-2 log-likelihood	11343.10	-5383.34	-1415.80	-2933.50	4711.80
ICC	0.414	0.371	0.490	0.075	0.537

Note. Continuation of previous table. RV = responses visited, HS = horizontal saccades, MTDI = matrix time distribution index, MTDSD = matrix time distribution standard deviation.

Multi-level Analyses of Eye Movements. Table 21 summarizes the results of a series of multi-level models that explore the effects of person ability, item difficulty, and rule knowledge on various eye movement variables. For the sake of briefness we present only the final best fitting models that resulted from a process of successively adding parameters (see supplementary data for all models). Some models estimated slope variances of zero or close to zero, in which case we present fixed slope models. Whenever random slopes seemed justified by the data we also tested if interaction effects could improve model fit. The predictor variables ability and difficulty were standardized to z-scores for easy comparability and interpretability. Rule knowledge was coded with zero (pretest) and one (posttest).

The first column in Table 21 shows results of a multi-level model for the Toggle Rate (TR) variable which reflects the frequency of saccades that ran between the two main areas of interest. It appears that TR was affected by person ability and by rule knowledge, but not by item difficulty. On the other hand, the absolute number of toggles (NTOG) was affected by item difficulty, but not by person ability. Thus, the reduction of TR due to higher ability was mainly driven by an increase of the processing time, while the absolute number of toggles remained constant across ability levels.

The relative time on matrix (RTM) reflects the cumulative timespan of all fixations within the bounds of the matrix area relative to total response time. This variable showed similar covariance patterns as TR and was influenced by ability but not difficulty. By breaking the RTM down to its numerator and denominator, we got clearer picture about how eye fixation time is allocated to the areas of interest. Difficulty increased both, the time on matrix (TM) and the time on alternatives (TA) which seems to mirror the overall effect of difficulty on processing time. In contrast, person ability did only increase the TM. In addition there was a significant interaction between the effects of ability and difficulty on the TM, which was also larger than the corresponding interaction effect on the TA. Cross referencing this with the pattern of results for response time shows that the increasing time spent on harder items by high ability subjects is mainly devoted to the matrix area, which is indicative of a constructive matching strategy.

Interestingly, this pattern was different for the relative time of the first response fixation (RFRF). This variable reflects the amount of time that was initially spent on the matrix area as a percentage of total response time and is based on the timestamp of the first fixation in any part of the response area. This variable was affected by item difficulty but not by person ability. Breaking this variable down to its numerator and denominator, reveals basically the same pattern as we found for overall response time, meaning that difficulty and ability both increase response time before and after the first response fixation. A closer look at the magnitude of effects, suggests that difficulty affected the absolute time after the first response fixation more (β = 12.051) than the time before (β = 5.017). The number of response alternatives visited (RV) showed the same pattern and was also affected by difficulty but not ability. Overall it seems like there were two distinct kinds of variables: Those affected by item difficulty (NTOG, RFRF, and RV) and those affected by person ability (TR and RTM).

Besides the differential effects of ability and difficulty on eye movement behavior, there was a pervasive effect of rule knowledge on all variables. This is consistent with the hypothesis that rule knowledge changes the strategic approach of problem solvers. Looking at the effects of rule knowledge on the decomposed eye movement variables revealed further details about how rule knowledge acts to change strategies. While ability reduced TR via an increase in time and constant NTOG, rule knowledge actually reduced NTOG while having no effect on response time. Thus, the effect of rule knowledge is more like the inverse effect of item difficulty. Similarly, rule knowledge had no significant impact on the TM, again acting unlike ability. The effect on RTM was rather driven by a weak negative effect on the TA and this pattern is also more comparable to an inverse effect of item difficulty. The effect of rule knowledge on RV was also like an inverse effect of difficulty, but the pattern of results for RFRF and its constituents was rather unique. Rule knowledge significantly increased the BFRF (β = 1.591) without affecting AFRF while difficulty was increasing AFRF stronger than BFRF.

Finally, the matrix time distribution index (MTDI), as originally proposed by Vigneau et al. (2006) was somewhat correlated with ability replicating their finding. Vigneau et al.'s interpretation was that "high-scoring individuals [were] trying to take into account all the information provided by each cell and low-scoring individuals [were] biasing their inspection mainly toward adjacent cells (last row and last column)" (p. 271). However, we noticed in our data that the MTDI distribution spanned rather evenly into positive and negative values (*min* = -1, *max* = .69), so a positive correlation should mean that, with higher ability, fixations were slightly biased towards the four top left cells of the matrix area. In order to test whether ability or difficulty affect the degree to which matrix time is distributed evenly versus unevenly it seemed more appropriate to analyze the absolute value of the MTDI. Positive values on this new variable should mean that matrix time is biased to either side, while values near zero suggest an even distribution. Since this variable was heavily skewed towards zero, we log-transformed the variable before the analysis. After this transformation the Abs(MTDI) was mostly affected by difficulty in that harder items produced an even distribution (β = -0.027). Note, the rather low ICC = .075 indicated that this variable was not as much affected by person level variance as other variables in this experiment (Table 21). Seeing that this variable might be hard to interpret, we conceived of an alternative and probably more direct way to index uneven time distribution among matrix cells. We took the sum of the duration of all fixations on each cell and calculated the standard deviation. We excluded cell nine from this statistic, since there should have been no information to glean from and there were consistently lower fixation times. An even distribution would result in a matrix time distribution standard deviation (MTDSD) near zero, while biased information seeking would result in increasing values. Again, this variable was heavily skewed towards zero, so we log-transformed the variable before the analysis. This variable was affected by difficulty (β = 0.399) in a way that it biased matrix cell fixations towards an uneven distribution. The MTDSD was also related to ability in the same way (β = 0.103). This finding runs contrary to Vigneau et al.'s interpretation of the original MTDI. Furthermore, two interaction effects suggest that ability increased bias even more with rule knowledge (β = 0.031) and that higher ability increased bias even more with higher difficulty (β = 0.117). In sum, three different indicators of matrix time distribution yielded three different, and partly contradicting, covariance patterns.

General Discussion

Are matrix reasoning strategies influenced by rule knowledge? How is response time related to ability, difficulty, and the use of strategy? The results from the current study offer some answers to these questions. The pilot study explored the feasibility of a within subjects paradigm. The results generally supported the hypothesis that eye movement behavior was affected by rule knowledge in a way that increases constructive matching. But we also noted that the two item blocks were not balanced and potentially confounded the results. After a careful revision of the item composition of the two blocks, we used the revised material in the main experiment. Here we also assessed participants' general mental ability with three additional measures of fluid intelligence.

An extensive series of multi-level models revealed some new insights into the relationship between ability, difficulty, response time, and eye movement behavior. We confirmed two of the hypothesized effects of rule knowledge: test performance increased, and eye movement behavior changed. We did not observe any of the hypothesized differences in the pattern of correlations between pretest and posttest conditions. That is, performance on the Raven test did not correlate differently with measures of fluid intelligence or with the backward digit span task as a measure of working memory capacity. We did, however observe a significant decrease of the correlation between Raven test scores and test time from pretest to posttest. Further multi-level analyses revealed that this effect might have occurred due to low ability subjects' increase in test time due to rule knowledge. Furthermore, the interaction effect of ability and difficulty on item response time suggests that persons with high ability somehow adapted to the difficulty of very hard items. Finally, by incorporating the effects on time into the effects on strategy we got a clearer interpretation of eye movement variables.

Eye Movement and Time

Response time on the Raven test was heavily influenced by item difficulty. Parallel to this finding, difficulty affected a distinct pattern of eye movement indicators. Harder items resulted in a higher absolute number of toggles (NTOG), in more response alternatives being taken into consideration (RV), and in an increased time after the initial inspection of the matrix (AFRF). That is, these behavioral variables are probably determined by item complexity, which generally indicates the amount of information that can be extracted from the item and how hard it is to extract this information. Many studies have shown that item complexity is a major determinant of item difficulty and processing time (Bethell-Fox et al., 1984; Green & Kluever, 1992; Mulholland et al., 1980; Primi, 2002) so this is nothing new, but the eye movement indicators give some insight into the way subjects spend their increased response times. Apparently they do a lot of comparisons between the problem matrix and the response alternatives and take more response alternatives into consideration. The lower RFRF on high difficulty items can, at least in part, be explained by increased time. Although difficulty affected processing time overall, the time after the first response fixations seems stronger affected than the time before. By definition, there can be no toggles or responses visited before the first response fixation, so the increased time after the first response fixation goes hand in hand with an increase of RV and NTOG. These findings suggest that difficulty increases the use of response elimination strategy and are well in line with the idea that it is a fallback strategy on harder items (Snow, 1980).

On the other hand, the overall mean RFRF was 48%, thus on average, subjects spent almost the entire first half of their total response time exclusively on the matrix area. By definition, there can be no response elimination during this time, thus constructive matching was the default strategy by the majority of subjects in this sample. This is in line with the mental model theory of reasoning which assumes that reasoners build an iconic and symbolic model of the elements in the problem and their interrelations (Johnson-Laird, 2004). From this model it is possible to deduce possible response alternatives and it should be only then, that reasoners consult response alternatives for comparison. It is possible that reasoners obtain disconfirming evidence from the response alternatives, for example, when they encounter multiple response alternatives compatible with their mental model. In that case they should try to make adjustments to the model based on additional information from the matrix area. This should be an iterative process that repeats until a satisfactory solution is found. However, if no solution is found then subjects have to revert to an alternative strategy that likely involves guessing. Previous descriptions of the response elimination strategy were rather elliptical and it is unclear how this strategy operates and when it is triggered. Given our assumption that constructive matching was the default strategy for a vast majority in our sample, it appears unlikely that participants eliminate responses from the very start of the reasoning process. We propose that each toggle down to the response area marks the completion of a model building or model updating process. The initial model usually takes the largest amount of time, almost half of the total response time. If disconfirming evidence is encountered during the first inspection of response alternatives then the initial model is updated and further information is sought from the matrix area. In this regard, TR or RV are not so much indicators of a distinct strategy, but indicators of iterations in a mental model building and updating process when the model does not lead to one definite solution.

Additionally, the RFRF was one of the main predictors of item performance and seems a crucial indicator for the initial mental model building process. Consider, for example, a study by Oberauer, Weidenfeld, and Hörnig (2006) in which participants were presented with a series of sentences describing spatial relationships between figural elements. Of the four sentences, the first sentence was by far the one with the longest reading time. Thus, it appears that setting up an initial mental model, even if it contains just two elements and one relation, takes the longest amount of time. Additional information is subsequently integrated into the initial model and takes less time.

Difficulty did also affect the MTDSD and made the distribution of fixations within the matrix more uneven. It is, however, not possible to say in what way it makes the distribution uneven since this statistic treats all the cells evenly. We surmise that biases are actually contingent on the content of a matrix. Raven items can differ greatly, for example, in the direction that the rules work or the focus that rules have on individual cells. For example, a one-of-each rule could evoke an even distribution of fixations within a row, while a plus rule might evoke fixations on the two addends in the first two cells of a row. Hence, further research is necessary to clarify the role of fixation distributions among matrix cells by controlling for matrix content. Meanwhile, Hayes et al. (2011) offered one solution by suggesting that eye movements of well performing subjects run systematically from cell to cell and row by row within the matrix area.

On the subject level, we found two variables that were related to ability instead of difficulty. The RTM was increasing and the TR was decreasing with higher ability. This finding is compatible and explainable with findings from the analysis of time. The lower TR due to ability was actually a result of increased response time while the absolute number of toggles remained rather constant. The increased RTM was also a result of increased response time, but especially due to increased response time on the matrix, while response time on alternatives was unaffected by ability. Thus, these two eye movement indicators give a good picture about how high ability subjects used their increased response time: They spend it almost exclusively on the matrix area while toggles between matrix and response alternatives remain constant. There was also a significant cross-level interaction on TM, meaning that on harder items, subjects with high ability seek even more information from the matrix. This might be evidence against the interpretation of response elimination as a fallback strategy, because then participants should direct less attention towards the matrix.

Bringing these results together it seems like person ability and item difficulty have somewhat differing effects on the information seeking pattern of test-takers. While ability increases the time invested and biases information seeking towards the matrix area, difficulty increases the frequency of information seeking from response alternatives.

Eye Movement and Rule Knowledge

All eye movement variables were affected by rule knowledge, generally supporting the idea that rule knowledge changes strategy. Looking at the effects of rule knowledge on the decomposed eye movement variables revealed further details about how rule knowledge acts to change strategies. Overall, the effect of rule knowledge on eye movements seems parallel to the inverse effect of difficulty.

The rule knowledge effect on RFRF was driven by an increased processing time of the matrix area and rather constant time after the FRF. This was unlike the effects of person ability or item difficulty, which were mostly reflecting the overall response time effect. This is clear evidence that subjects were trying harder to figure out how to apply the learned rules to the problem in the very beginning of the problem solving process. That is, rule knowledge stimulated subjects to spend more time on building an initial mental model. The fact that rule knowledge did also reduce the number of toggles and the number of responses visited provides evidence that the constructive matching process was overall more successful, leading to a response in fewer updating iterations.

A unique effect of rule knowledge was present for HS and this was likely due to the fact that the rule teaching video explained the rules row-wise. If the rules were taught vertically we would likely have observed an increase of vertical saccades, but this should not have affected any of the other variables as they are defined independent of gaze direction within the matrix area.

In sum, rule knowledge does not increase response time while reducing the number of toggles, reducing the amount of responses visited, reducing the absolute time on alternatives, and increasing the initial processing time of the matrix. This pattern of results does not only indicate a frequent usage of a constructive matching strategy, but also indicates that the strategy is more successful when the rules are known.

Time

Our first approach to the analysis of response time data was a conjoint item response and time model (CIRT), which estimates parameters for time and performance on the item level and the person level. The results replicated previous findings and encouraged the notion that item level and person level effects should be analyzed separately. We found that persons with higher ability were overall slower and that items with increasing difficulty made item response times slower. We also found that response latencies were correlated across different tasks of general fluid ability, thus giving further evidence for the idea that there is a general test-taking time factor (Roberts & Stankov, 1999).

Dodonova and Dodonov (2013) raised concerns about the CIRT approach because “a single estimate of the speed-accuracy correlation at the person level can hide the true relations between these variables, which are likely to vary as a result of varying item difficulty” (p. 9). We were able to confirm a varying effect of ability on response time, depending on item difficulty in a multi-level analysis. That is, subjects with high ability were as fast as the rest on easy items (or slightly faster even) but seemed to adapt to increasing item difficulty with longer response times. Subjects with low ability were also adapting and increasing their response times to difficulty but at a significantly lower rate. Although not reported here, we also found similar patterns of results with measures of fluid intelligence other than the Raven (see supplementary material). This explains the, at a first glance, contrary effects of item difficulty and person ability. That is, on any given item, difficulty increases response time and easiness reduces response time. That should mean that a person with high ability, for whom the item is easy, should be faster than a person for whom the item is hard. But the person with high ability will invest even more time to solve the hard items and thus achieve a higher probability for a correct response. That is why there is generally a positive correlation on the person level.

Sternberg made the point: “… not that more intelligent people are inherently slower than less intelligent ones, but rather, that the key to their intelligence insofar as speed of processing is concerned is not just that they are fast or slow, but that they know when to be which” (Sternberg, 1986, p. 269). But how do they know? At least two explanations seem plausible, both of which suggest a role of motivation. First, the positive correlation might be an epiphenomenon of motivation as a common underlying cause. That is, motivation would affect effort that is related to the investment of cognitive resources. Assuming that the capacity of cognitive resources is limited by general ability and independent of motivation, then higher motivation would affect the timespan over which cognitive resources are allocated to the task of interest. This in turn might increase the probability of finding a correct solution to a given problem. As such, motivation would be an independent factor of general ability but also important in explaining test results. The challenging question in this context is, whether the low ability subjects would have scored higher had they taken themselves more time?

A second plausible explanation assumes that motivation or effort, as reflected by time investment, is a result of test experience. As a test-taker progresses through the items of a test they may experience how hard or easy the test is for them (Mitchum & Kelley, 2010), which would in turn be dependent on the individual's general ability. This might in turn affect their motivation as pictured by Weiner's attributional theory of motivation (Weiner, 1985). That is, if an individual experiences the test items as challenging then motivation might decline thus items are not given the proper mental resources. On the other hand, if an individual feels confident in his answers, then this might lead them to accept the challenges of very hard items and invest a lot of time an effort into the solution process.

The distinction is similar to the one between a trait and a state, which means that actually both explanations can be true at the same time and cannot be answered conclusively with the current data. Nonetheless, the finding that rule knowledge did improve test performance and strategy but not response times, might be indicative for the first explanation, i.e. that motivation acts as trait and has an effect independent of ability and test experience. Also, the CIRT estimated traits of person speed correlated across different fluid intelligence tasks and provide evidence for an independent motivational cause. We also found that the correlation between time and performance was significantly lower and almost zero after rule knowledge and this effect was related to the finding that subjects with low ability increased their response times with rule knowledge but high ability subjects did not. This could mean that subjects with low ability did profit especially from the rule teaching intervention but that was not the case, as there was no significant interaction between ability and the intervention effect on test performance (see supplementary material). So all subjects did profit from knowing the rules to about the same degree, but low ability subjects did also increase their response times and this could have a motivational cause.

Further research is necessary, that would incorporate experimental manipulations of the time limits or manipulations of the motivational state via incentives. One could, for example, limit the response time windows for high ability subjects to the same level as those of low ability subjects and see if performance would drop to about the same level, or one could persuade low ability subjects to take as much time as high ability subjects. There are, of course, methodological challenges with the implementation of such designs because there is no uniquely typical response time span for all subjects with certain ability. Setting a time limit to some average would not work for many individuals. Maybe, the estimation of such time limits could be informed by an on-line evaluation of eye-movements. That is, if it was possible to identify distinct stages of reasoning in eye movements then it could be possible to set time limits or provide incentives based on this information. The current research can be seen as a first step towards the identification of such stages in matrix reasoning.

Stable Correlations

The hypothesis concerning changes in the correlations with measures of fluid intelligence was not confirmed in this sample. This parallels previous findings in which correlations with creativity or measures of rule discovery did not change due to rule knowledge (Loesche et al., 2015). Thus, changes in strategy could explain previous findings in which the correlation with working memory capacity raised with rule knowledge. Constructing a mental model of the problem and deducting possible response alternatives is arguably demanding to working memory capacity because all the elements and relations need maintained during reasoning. Eye movement indicators revealed that rule knowledge increased the time of the initial model building phase, as indicated by a delayed first response fixation. This leads us to suggest that the knowledge of possible relations is already incorporated into early mental models. This did also result in a more successful reasoning and less frequent model updates, as indicated by lesser toggles between the matrix and the response alternatives.

Our original idea was that working memory is not crucial for generating possible rules during the reasoning process because current theories about working memory are mostly concerned with processing and maintenance of already available information. Thus, other processes should be relevant, processes that control the emergence of new information. So far there was no other process that seems empirically plausible for the explanation of rule generation, so how do people come up with the rules? Most of the rules can probably be considered as abstractions of more or less common knowledge (e.g., addition or subtraction). So rules probably need not be created and induced from scratch, but abstract concepts of the rules need to be sought out from conceptual knowledge in long term memory (Hunt, 1974).

Arguably in our experiments, people did not learn completely new rules. Everybody should be aware of concepts such as addition, or the even distribution of things, or the progressive change of attributes. So the rule teaching videos did not exactly teach something new but were rather activating certain concepts and made them more accessible. It is questionable if the teaching of rules would have had the same success in populations where these common concepts are unknown, like very young children for example. Intelligence test items for preschool or early elementary school do often contain figural reasoning tasks but the underlying rules are much simpler than in adult versions. Arguably, adult rule concepts would not work for young children. It seems unlikely that preschool children could induce new complex concepts all on their own during an IQ test when there is no prior knowledge. Thus, reasoning would depend on knowledge. Complementary to investment theory (Cattell, 1963; Schweizer & Koch, 2001), this is probably one reason why there is a relation between fluid and crystallized intelligence and it is a possible explanation for the Flynn effect (Armstrong & Woodley, 2014; M. C. Fox & Mitchum, 2013).

Intriguing questions remain unanswered as to under what circumstances certain concepts get activated. For example, it might not only be relevant that conceptual knowledge of a rule is present at all, but also the accessibility of said knowledge. Research from Rosen and Engle (1997) suggests that working memory capacity is a limiting factor in a verbal fluency task that measures the accessibility of long term memory content. Working memory is supposedly responsible for monitoring and enables a controlled strategic search of accessible knowledge. Another theory proposes the link to long term memory works through the episodic buffer, and is separate from the working memory system (Baddeley, 2000). Baddeley thought of this memory component as being episodic in nature, however abstract concepts underlying figural reasoning tasks are rather semantic. More relevant then is probably the visuospatial sketchpad which Baddeley supposed to be relevant for semantic content. The episodic buffer, to date, is not very well understood, both theoretically and empirically, and has often been used to explain puzzling results (Baddeley, Allen, & Hitch, 2010). Thus, further research is necessary to investigate how mental models in working memory match up with conceptual knowledge about abstract rules in long term memory.

Limitations

Contrary to previous results (Loesche et al., 2015) the correlation with a backward digit span did not increase with rule knowledge in the current sample. The focus of this study was on the relationship to gf, so there was only a single task to measure working memory capacity and this was possibly not fully capturing the construct. Thus, possible effects could have been overshadowed by task specific sources of variance. Additionally, the effect of rule knowledge was comparably weak in the current sample (about 5% more items solved with rule knowledge) than in previous between-subjects designs (about 10% more items solved with rule knowledge; Loesche et al., 2015). Reasons could be due to the alignment of instructions between pretest and posttest or possibly due to the limited ability range in the current sample. Compared to middle school students in previous research, the current sample comprised of undergraduates who arguably were a selective subpopulation with relatively higher and fully developed working memory capacity. This might have factored into the observed correlations and would suggest a developmentally differential role of working memory capacity for the application of certain strategies. Despite these possible explanations, this raises doubt on our previous finding and further research is necessary to investigate the robustness of this effect. It seems at least plausible to expect higher reliance on working memory because the observed strategy shifts should demand more working memory capacity.

A further limitation is that there was no true control group, which could have been helpful in clarifying the interaction effects due to unequal item blocks. However, the whole purpose of the within-subjects design was to avoid the necessity of such a control group and maximize the test-power in a limited sample size. Additionally, the fact that we already found strong effects on eye movement behavior in a between-subjects design (Loesche et al., 2015), strengthens the interpretation that the current findings were attributable to rule knowledge.

Finally, short item blocks of fourteen items each in the main experiment, might have caused estimation problems in some of our multi-level models. Although this should not have affected the central pattern of results, it could be useful in future studies to utilize automatic item generators (e.g., Arendasy & Sommer, 2005; Matzen et al., 2010) to create fully parallel items in large quantities.

Conclusion

Rule knowledge affects strategy for the better. We interpreted our results in light of the model theory of reasoning (Johnson-Laird, 2005). Eye movements suggest that rule knowledge affects the initial encoding and mental model building process and leads to a response with fewer iterations. That is, mental models are more comprehensive and this can explain why previous studies have found an increased correlation with measures of working memory in samples where the capacity was not yet full-grown (Loesche et al., 2015).

For decades, researchers have assumed that two strategies can be applied in figural reasoning. Seeing that eye-movement variables were largely affected by item difficulty it seems questionable to assume qualitative strategic differences on the person level. Our interpretation is rather, that differences in cognitive ability make differences in the comprehensiveness of mental models. Response elimination as a purely backwards reasoning strategy does probably not exist and would be incompatible with the idea, that people build mental models during reasoning. We maintain that eye-movement indicators that were previously connected to a response elimination strategy (e.g., toggle rate) are better interpreted as indicators of the efficiency of the model building process. We propose that response elimination would only occur after the initial constructive matching fails to yield a satisfying solution and is more guessing than strategy.

Last, but not least, our analyses of response times suggest that subjects of high ability do generally employ more effort in their reasoning and even more so on challenging problems. This hints towards a motivational aspect in reasoning that is, at the same time, closely linked to ability. Recent discussions about the application of speeded versus power tests (Ackerman & Ellingsen, 2016) could be informed by this finding, suggesting that high ability individuals can only play their cards right when they have the proper time for it.

Epilogue

The first part of this dissertation gave an overview of the possibilities of cognitive assessment in infants and children. One major finding in this field of research is that with younger age, a reliable assessment is increasingly hard to achieve. Methods that attempt to assess cognitive abilities in infants and children suffer from low reliability and it is harder to find evidence for a psychometric g-factor in very young children than in adults. A theoretical account by van der Maas et al. (2006) suggested that the g-factor is an epiphenomenon that emerges during cognitive development in early years from the interplay of cognitive processes. Hence, research in psychological assessment of talent and giftedness could gain new insight by focusing on well described cognitive functions. Recent studies have shown that working memory is a good predictor of early school achievement (Alloway & Alloway, 2010; Fischbach, Preßler, & Hasselhorn, 2012). Furthermore, working memory has gained much attention in intelligence research over the past two decades (e.g., Ackerman et al., 2005). That is why the role of working memory capacity for measures of intelligence was of particular interest in the empirical parts of this dissertation.

The experiments in Part 2 were inspired by the idea that rule induction would not depend on working memory capacity. The first three experiments explored the correlational pattern of the Raven test with other measures, under the condition of known rules and under normal testing conditions. The finding that the correlation with working memory capacity increases under the condition of known rules suggests that rule induction is not dependent on working memory. However, testing the inverse rationale of this hypothesis did not reveal the expected results. The idea was that pure measures of rule induction should correlate less with the Raven test under the condition of known rules. But the Brixon Rule Anticipation test (see Crescentini et al., 2011) and creativity measures correlated the same with matrix reasoning under both experimental conditions. Experiment 4 additionally revealed that eye movement patterns were changing as a result of the experimental manipulation of rule knowledge. The results suggest that matrix reasoning with known rules involves less frequent saccades between the matrix and response alternatives and longer inspection times of the matrix. In combination with previous research (Vigneau et al., 2006), this pattern of results indicates that reasoners deploy a more advanced and effective strategy that is usually deployed by individuals with high ability. Assuming that the advanced solution strategy is working memory demanding, this finding offers an alternative explanation for the finding of increased working memory correlations. That is to say, correlations with working memory measures might have increased because knowing the rules enabled reasoners to engage in more working memory demanding strategies.

The aim of the experiments in Part 3 was to investigate this hypothesis further. The basic idea was to replicate the results of Experiment 4 in a larger sample and to utilize a within-subjects design for greater test power. Indeed, the results replicated the finding of changing eye movement patterns due to rule knowledge and provide further insight into the problem solving process via the consideration of response latencies. The data does generally support the notion that test-takers are more likely to engage in a constructive matching strategy. This strategy involves the mental construction of a possible answer to then match it against the given response alternatives. This strategy was interpreted against the background of mental models theory, which assumes that reasoning involves the construction of mental models that form abstractions of perceived stimuli (Johnson-Laird, 2004, 2005). According to this interpretation, there should be no qualitative inter-individual differences in solution strategies. The data revealed that eye-movements were predominantly changing within subjects as a function of task difficulty, suggesting that variables that were previously interpreted to reflect a qualitatively different strategy are more likely to reflect the progress of the mental model building process. I will elaborate more on the implications of these results for inductive reasoning in the next section and will then conclude with thoughts on intelligence.

Thoughts on Inductive Reasoning

After reviewing a host of research on reasoning and analyzing first-hand empirical results in the current dissertation, I have come to revise my initial theory about inductive reasoning. I first thought that inductive reasoning, as measured by matrix reasoning items, can be described by two components that were inspired by Carpenter et al. (1990). One must be chiefly influenced by working memory capacity. This was a necessary assumption, since measures of working memory capacity have reliably proven to share a great amount of variance with measures fluid intelligence in general, and with matrix reasoning in particular. Initially, working memory was thought to be connected to goal management as hypothesized by Carpenter et al. (1990). According to this notion, working memory would be a requirement for active maintenance and updating of task goals during the execution of various reasoning processes. But the concept of goal management is probably more relevant when researchers are trying to model human cognition in artificial cognitive architectures where they provide a guiding structure (e.g., Choi, 2011). Computer programs are usually systematic so that the same input leads to the same output. However, that is not necessarily the case in natural human cognition. The discussion of results in Part 3 of this dissertation concludes that a mental model theory (Johnson-Laird, 2004, 2005) can explain why matrix reasoning relies on temporary storage and processing. Supported by observations of eye-movements and response times, it was conjectured that a central part of matrix reasoning consists of the consecutive build and update of an abstract mental representation of the matrix problem. This implies that working memory capacity would be completely occupied with a mental model that is characterized by figural elements (or abstractions thereof) and their relations.

The relations are central to the second supposed process, namely rule induction. It is evident that John Raven created his matrices with some rules in mind, and Carpenter et al. (1990) empirically showed that people are able to recognize and verbalize these rules. They listed five rules that were based on a small sample of think-aloud protocols. However, these were only the ones that were actually recognized and verbalized in their samples of 12 and 22 students. There are certainly more rules involved in Raven’s APM or other matrix reasoning tests and this is why some items could not be classified by Carpenter et al. (1990).

Regardless of the actual amount of rules, one central research question of this dissertation was how people can recognize these rules. Based on theoretical considerations, it was initially hypothesized that working memory cannot be responsible for rule induction. The logic was that rules need to be generated, so the cognitive process responsible for this needed some productive feature. Thus, one of the empirical approaches towards this process was the consideration of creativity, which was unsuccessful. Creativity turned out to be a particularly difficult construct. There are ways of measuring it (Silvia et al., 2008), but when it comes to the practical implementation, there are so many points that require a decision of some sort, that in the end one can barely speak of an objective and reliable measurement. There is, for example, the decision about what responses should be excluded for being beside the point or unrealistic, and there are multiple options for the degree to which guidelines confine ratings. Furthermore, the inter-rater agreement was alarmingly low (see Part 2) which was possibly a result of providing sparse confinement. However, more rigorous guidelines could have compromised ratings in a way that they reflect the guidelines instead of actual ratings, and guidelines are subjective as well. The bottom line is that creativity, in terms of productive and divergent thinking, is hardly measurable and it requires great time and effort to even try.

Another approach to make the underlying processes directly detectable was based on a method from research in cognitive neuroscience. The Brixton Spatial Anticipation task was shown to trigger activation in brain areas that were different, depending on whether the current phase in the task required search for a rule or application of a rule (Crescentini et al., 2011). However, research presented here suggests that performance on this task was significantly correlated with working memory capacity. In retrospect, this is not surprising because the task has a storage demanding component. Participants were asked to track the position of a circle on a 12x2 spatial array of slots. This circle was jumping from slot to slot, following a hidden rule. In order to recognize the pattern, it was inevitable to memorize at least two consecutive slots and to process whether the current guess is right or wrong. While the recognition of hidden rules certainly has a part in this task, the reliance on working memory is just too big to consider it a pure measure of rule induction.

Finally, one of the theoretical assumptions put forward in Part 3, was that rule induction must be relevant in other inductive reasoning tasks that are common in the field of intelligence research. To cut a long story short, there were again none of the predicted patterns in the results. Correlations were not changing as a function of rule knowledge. However, the results from the observation of eye movements in the final experiment in Part 2 and in the experiments in Part 3 suggest that previous results, that were thought to be connected to rule induction, could be related to the use of certain reasoning strategies. As pointed out in the beginning of this paragraph, this points towards a mental model theory of reasoning.

The conclusions to be drawn empirically about rule induction are limited, mostly because none of the approaches to find a pure measure were particularly successful. However, I can try to advance our understanding of rule induction and inductive reasoning theoretically. One key revelation came during the discussion of the results in Part 3: The manipulation of rule knowledge, the main characteristic of the paradigms in all experiments here, did not manipulate rule knowledge. The idea is that all the rules that were taught in the current experiments (plus, minus, one of each, constant, progress) were common knowledge in the population underlying the samples here. In other words, the rules are assumed to be already present as abstract concepts in long-term memory of most human beings with a certain age and education. Unlike my first assumption, rules need not at all be created from scratch. Instead, the representations of said rules need to be retrieved from long-term memory and correctly matched to the current mental model. There is no need for a rule induction process that generates these rules and makes them a conscious representation in working memory. The process is probably more like a thorough search for matching representations in long-term memory, similar to the search for the name to a familiar face. This might be quite easy but can prove to be quite troublesome, as research on the tip-of-the-tongue effect has shown (Burke, MacKay, Worthley, & Wade, 1991).

A long-term-memory-theory of rule induction would assume that teaching of rules to a matrix reasoning task pre-activates the corresponding representations in long-term memory. As a result, more time and effort can be devoted to building the mental model instead of trying to match rules. Verguts et al. (1999) proposed that rule generation works similar to sampling from an urn, and this seems like a very appropriate analogy to the rule induction process if one assumes that the urn is filled with knowledge. Long-term memory can be thought of as an associative network (Raaijmakers & Shiffrin, 1981) and each representation in long-term memory has to have a certain pattern of neuronal activation or neural structure (Kandel, 2001; Moscovitch et al., 2005). On the other hand, mental models should correspond to certain neuronal activation patterns in the frontal cortex and visual cortex (Linden, 2007). There has to be a binding entity between working memory and long-term memory that would (1) guide the search through long-term memory and (2) give some sort of signal if a match is found.

The search might be guided by the associative nature of long-term memory itself, meaning that one activation leads to the next and spreads out. In this regard, creativity could be reconsidered, not so much as a measure of productive thinking as presented here, but rather as a measure of the associative hierarchy of knowledge in long-term memory (Benedek, Könen, & Neubauer, 2012; Mednick, 1962). The signal could be grounded in an affective reaction, similar to the one being assumed to occur with problem solving insight (Topolinski & Reber, 2010). The strength of that signal could be determined by the degree of overlap between neural activation in working memory and neural structure in long-term memory. It is possible that people find a rule concept in long-term memory that kind of matches, but is not entirely correct. There could still be some discrepancy so that the rule cannot fully explain everything, or there might even be some contradicting aspects in the problem, and that is what people can feel. The idea is related to Festinger's theory of cognitive dissonance (Festinger, 1957). Although Festinger did not define the theory on the basis of a distinction between working memory and long-term memory, the basic idea should hold. Take the famous example of a smoker who learns that smoking is unhealthy. The habit of smoking should be connected to various representations in long term memory. Especially episodic memory should contain various episodes where smoking was a part. If a smoker is confronted with evidence for the unhealthiness of smoking, this information is processed as ongoing cognition in working memory and embodies a discrepancy to long-term memory representations. According to Festinger's theory, people would feel uneasy about such cognitive dissonance and would try to resolve the issue. Such affective states might in turn be the underlying cause of confidence judgements in matrix reasoning that can be either over-confident or under-confident (Mitchum & Kelley, 2010). Under-confidence could be the result of said discrepancies when long-term memory concepts do not fully correspond to mental models. Over-confidence could be the result of inaccurate or incomplete mental models, to which some long-term memory content matches subjectively well, but is actually inadequate. "If high and low performers differ in their mental representations of items, this could also affect how perceptual and item features are used as cues for confidence monitoring" (Mitchum & Kelley, 2010, p. 708).

The implication for inductive matrix reasoning is twofold. First, the success of reasoning is dependent on working memory capacity in that the mental model needs to capture as many features as possible from the problem space. Second, reasoning is dependent on the accessibility of relevant knowledge structures in long-term memory. Search in long-term memory might be directly controlled by working memory as suggested by Unsworth, Brewer, and Spillers (2013). Furthermore, Baddeley (2000) suggested that the link between working memory and long-term memory, the episodic buffer, is closely linked to working memory. This view implies that retrieval from long-term memory does not only depend on long-term memory itself, but also on working memory. This stresses how cognitive processes involved in inductive reasoning do not work additively, but interact and influence each other multiplicatively.

Thoughts on Intelligence

What is intelligence? That was the opening question of this dissertation and really one of the first questions I asked myself when I was looking for a topic. I was originally aiming to do research on talent and giftedness, but got stuck with this very fundamental question that many have asked before me. While the short answer would simply be “I still don’t know”, the long answer might cover at least some ground towards an understanding.

McGrew (2009) presented the CHC-theory of the psychometric structure of intelligence. This theory is based on previous g-factor theories and stresses the importance of *g*; however, it also showcases how diverse psychometric measurements of intelligence are. At the intermediate stratum II, there are at least ten broad ability domains: Fluid reasoning, comprehension knowledge, short-term memory, visual processing, auditory processing, long-term memory, processing speed, reaction and decision speed, reading and writing, and quantitative knowledge. All of them are thought to be chiefly accounted for by the g-factor. But what could explain individual differences in performance on simple reaction time tasks as well as performance on reading and writing tests? The key term here is: individual differences. If *g* really was some sort of “energy or power which serves in common the whole cortex” (Spearman, 1923, p. 5), then there has to be variation in its magnitude across individuals. However, given that we can already distinguish between so many abilities and have, in part, established different brain areas as their basis, there is little more left that *g* could be. Instead I favor the view that the mutualism model of intelligence proposes (van der Maas et al., 2006). According to this theory, the g-factor is just an epiphenomenon of the interplay of many distinct cognitive processes (see Part 1). I have argued in the previous section that working memory and long-term memory work together to enable inductive reasoning. Thus, individual differences can actually stem from either cognitive process.

Most of what is currently known about the human brain, suggests that it is compartmentalized. There is, for example, a part in the brain responsible for vision (visual cortex) and other parts (Broca’s and Wernicke’s area) for language (see Rosenzweig, Breedlove, & Watson, 2005). We already touched the distinction between long-term memory and short-term memory (Shiffrin & Atkinson, 1969). Thus, there is no measurable intelligence but only batteries of tests that serve to provide an index of a more or less representative sample of cognitive functions, meaning that intelligence is just what the intelligence test measures (van der Maas, Kan, & Borsboom, 2014). There is nothing wrong with an index and it has already proven to be very useful on various occasions (e.g., military recruiting, job applications). However, there are instances where such an approach has drawbacks.

For example, studies about the genetic origins and heritability of intelligence, are more often than not working with a g-factor definition of intelligence. Chabris et al. (2012) reported three studies that investigate the correlations between various single-nucleotide polymorphisms (SNPs) and general intelligence. All three studies measured *g* with a different test battery. Notably, Chabris et al. reported that 11 of the 12 investigated SNPs were not significantly correlated with general intelligence, although all of them have been reported in previous studies to be of significance. Chabris et al. argue that previously reported results were false positives from underpowered samples and discuss “that *g* is a highly polygenic trait on which common genetic variants individually have only small effects” (p. 1320). This must be especially true when *g* is being measured with a broad range of cognitive tests. Each gene on the DNA strand codes for certain proteins, the building blocks of living organisms, of which there are thousands (M.-S. Kim et al., 2014). It is hardly conceivable how single base-pairs in a gene, which only result in slight variations in the resulting proteins, can account for a broad amalgamation of cognitive abilities. Instead, focusing on the investigation of narrow cognitive functions might be more productive.

As a further example, in research on learning disabilities like dyslexia it is very common to look at general intelligence as a benchmark against which reading deficits are compared (Hasselhorn & Schuchardt, 2006). However, the cognitive deficit of dyslexia is likely a deficit of working memory (Brandenburg et al., 2015), and given that working memory is even an explicit part of some intelligence batteries, it appears to be more reasonable to compare the reading deficits to other cognitive functions that are independent of working memory.

Finally, applied psychological assessment could gain something by turning away from the g-factor. Looking back at all the research that was discussed in this dissertation, it seems to me that if there really was a single process responsible for *g*, then it would probably be working memory. There are some of the broad abilities in CHC-theory that share little variance with working memory capacity, like processing speed for example (Conway et al., 2002). Thus, working memory is probably not exactly like the omnipresent mental power that Spearman and other *g*-proponents have envisioned, but it is central to a broad range of tasks (Baddeley, 1986). Some studies have already shown that working memory measures are predictive for school achievement (Alloway & Alloway, 2010; Cowan et al., 2005; Fischbach et al., 2012; Vock & Holling, 2008) and it is only a matter of time until they are being used in job applications and military recruiting. This could have one major advantage compared to general intelligence measures: Making norms superfluous.

Intelligence tests need a norming sample to compare individual assessments against (relative measurement) because test scores on their own are meaningless and cannot be compared to scores from a different test. Creating such norms is a lot of work and needs to be repeated every so often to account for the Flynn effect (Flynn, 1984). Research on working memory has already made some advances towards a direct approximation of the underlying capacity which is estimated to range from three to five chunks (Cowan et al., 2005). Thus, instead of an IQ, it would theoretically be possible to measure human intellectual potential directly as a capacity, just like measuring running speed and jumping height in sports. Of course, there would still be need for a reliable instrument and in the end the choice for an instrument is a matter of consensus, just like a running distance of 100m is a consensus. However, in sports there is no need to compare performance against a representative sample of runners because the speed is meaningful on its own. Research on the Flynn effect might actually profit from this, because estimating this effect goes not without making some simplifications. For example, Flynn’s (1984) original estimate translates to 3 IQ points per decade but that can hardly mean that people today have an average IQ of 115 compared to 50 years ago. It is necessary to express the Flynn effect in terms of IQ points because its estimation is based on many different tests. The estimation of working memory capacity is, to a certain degree, as well dependent on the task that is being used, so there is still work left to be done before it is possible to obtain an absolute measurement. But once it is possible, this might revolutionize human cognitive assessment.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: the same or different constructs? *Psychological Bulletin, 131*(1), 30–60. doi: 10.1037/0033-2909.131.1.30

Ackerman, P. L., & Ellingsen, V. J. (2016). Speed and accuracy indicators of test performance under different instructional conditions: Intelligence correlates. *Intelligence, 56*, 1-9. doi: 10.1016/j.intell.2016.02.004

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-Practice Recommendations for Estimating Cross-Level Interaction Effects Using Multilevel Modeling. *Journal of Management, 39*(6), 1490-1528. doi: 10.1177/0149206313478188

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106*(1), 20-29. doi: 10.1016/j.jecp.2009.11.003

Arendasy, M. E., & Sommer, M. (2005). The effect of different types of perceptual manipulations on the dimensionality of automatically generated figural matrices. *Intelligence, 33*(3), 307–324. doi: 10.1016/j.intell.2005.02.002

Arendasy, M. E., & Sommer, M. (2013). Reducing response elimination strategies enhances the construct validity of figural matrices. *Intelligence, 41*(4), 234-243. doi: 10.1016/j.intell.2013.03.006

Armstrong, E. L., & Woodley, M. A. (2014). The rule-dependence model explains the commonalities between the Flynn effect and IQ gains via retesting. *Learning and Individual Differences, 29*, 41-49. doi: 10.1016/j.lindif.2013.10.009

Baddeley, A. (1986). *Working Memory*. Oxford, UK: Oxford University Press.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417-423. doi: 10.1016/S1364-6613(00)01538-2

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), 829–839. doi: 10.1038/nrn1201

Baddeley, A., Allen, R. J., & Hitch, G. J. (2010). Investigating the episodic buffer. *Psychologica Belgica, 50*(3-4), 223-243. doi: 10.5334/pb-50-3-4-223

Baddeley, A., & Hitch, G. (2007). Working memory: past, present ... and future? In N. Osaka, R. H. Logie, & M. D'Esposito (Eds.), *The Cognitive Neuroscience of Working Memory* (Reprinted 2008 ed.). New York: Oxford University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 48. doi: 10.18637/jss.v067.i01

Benedek, M., Könen, T., & Neubauer, A. C. (2012). Associative abilities underlying creativity. *Psychology of Aesthetics, Creativity, and the Arts, 6*(3), 273–281. doi: 10.1037/a0027059

Berg, C. A., & Sternberg, R. J. (1985). Response to novelty: Continuity versus discontinuity in the developmental course of intelligence. In W. R. Hayne (Ed.), *Advances in Child Development and Behavior* (Vol. 19, pp. 1-47). Orlando: JAI.

Berti, S. (2010). Arbeitsgedächtnis: Vergangenheit, Gegenwart und Zukunft eines theoretischen Konstruktes. *Psychologische Rundschau, 61*(1), 3-9. doi: 10.1026/0033-3042/a000004

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. doi: 10.1016/0160-2896(84)90009-6

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence, 103*(1–2), 347–356. doi: 10.1016/s0004-3702(98)00055-1

Bolker, B. M. (2015). Linear and generalized linear mixed models. In G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Eds.), *Ecological Statistics: Contemporary Theory and Application*. Oxford, UK: Oxford University Press.

Boring, E. G. (1923). Intelligence as the Tests Test It. *New Republic, 36*, 35-37.

Brandenburg, J., Klesczewski, J., Fischbach, A., Schuchardt, K., Büttner, G., & Hasselhorn, M. (2015). Working Memory in Children With Learning Disabilities in Reading Versus Spelling: Searching for Overlapping and Specific Cognitive Factors. *Journal of Learning Disabilities, 48*(6), 622-634. doi: 10.1177/0022219414521665

Büchner, P., & Krüger, H.-H. (1996). Soziale Ungleichheiten beim Bildungserwerb innerhalb und außerhalb der Schule : Ergebnisse einer empirischen Untersuchung in Hessen und Sachsen-Anhalt. *Aus Politik und Zeitgeschichte, 11*, 21-30.

Burgess, G. C., Gray, J. R., Conway, A. R. A., & Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology: General, 140*(4), 674-692. doi: 10.1037/a0024695

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language, 30*(5), 542-579. doi: 10.1016/0749-596X(91)90026-G

Cameron, J., Livson, N., & Bayley, N. (1967). Infant vocalizations and their relationship to mature intelligence. *Science, 157*(3786), 331-333. doi: 10.1126/science.157.3786.331

Carman, C. A., & Taylor, D. K. (2010). Socioeconomic status effects on using the Naglieri Nonverbal Ability Test (NNAT) to identify the gifted/talented. *Gifted Child Quarterly, 54*(2), 75-84. doi: 10.1177/0016986209355976

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), 404–431. doi: 10.1037/0033-295x.97.3.404

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

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

Chabris, C. F., Hebert, B. M., Benjamin, D. J., Beauchamp, J., Cesarini, D., van der Loos, M., . . . Laibson, D. (2012). Most Reported Genetic Associations With General Intelligence Are Probably False Positives. *Psychological Science, 23*(11), 1314-1323. doi: 10.1177/0956797611435528

Choi, D. (2011). Reactive goal management in a cognitive architecture. *Cognitive Systems Research, 12*(3–4), 293-308. doi: 10.1016/j.cogsys.2010.09.002

Chuderski, A., & Neecka, E. (2012). The contribution of working memory to fluid reasoning: Capacity, control, or both? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 38*(6), 1689-1710. doi: 10.1037/a0028465

Chuderski, A., Taraday, M., Nęcka, E., & Smoleń, T. (2012). Storage capacity explains fluid intelligence but executive control does not. *Intelligence, 40*(3), 278–295. doi: 10.1016/j.intell.2012.02.010

Colom, R., Abad, F. J., Quiroga, M. Á., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence, 36*(6), 584–606. doi: 10.1016/j.intell.2008.01.002

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence, 32*(3), 277–296. doi: 10.1016/j.intell.2003.12.002

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163-183. doi: 10.1016/s0160-2896(01)00096-4

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 769-786. doi: 10.3758/bf03196772

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42–100. doi: 10.1016/j.cogpsych.2004.12.001

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*(8), 1754-1768. doi: 10.3758/bf03195936

Crescentini, C., Seyed-Allaei, S., De Pisapia, N., Jovicich, J., Amati, D., & Shallice, T. (2011). Mechanisms of rule acquisition and rule following in inductive reasoning. *The Journal of Neuroscience, 31*(21), 7763-7774. doi: 10.1523/jneurosci.4579-10.2011

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450–466. doi: 10.1016/s0022-5371(80)90312-6

Dartnall, T. (2002). *Creativity, cognition, and knowledge - An interaction*. Westport, CT: Praeger.

Dodonova, Y. A., & Dodonov, Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. *Intelligence, 41*(1), 1-10. doi: 10.1016/j.intell.2012.10.003

Doerfler, T., & Hornke, L. F. (2010). Working style and extraversion: New insights into the nature of item response latencies in computer-based intelligence testing. *Journal of Research in Personality, 44*(1), 159-162. doi: 10.1016/j.jrp.2009.12.002

Ekstrom, R. B., French, J. W., Harman, H. H., & Diran, D. (1976). *Kit of Factor-Referenced Cognitive Tests*. Princeton, NJ: Educational Testing Service.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380–396. doi: 10.1037/1082-989x.3.3.380

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309–331. doi: 10.1037/0096-3445.128.3.309

Fagan, J. F., Holland, C. R., & Wheeler, K. (2007). The prediction, from infancy, of adult IQ and achievement. *Intelligence, 35*(3), 225-231. doi: 10.1016/j.intell.2006.07.007

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Fischbach, A., Preßler, A.-L., & Hasselhorn, M. (2012). Die prognostische Validität der AGTB 5-12 für den Erwerb von Schriftsprache und Mathematik. In M. Hasselhorn & C. Zoelch (Eds.), *Funktionsdiagnostik des Arbeitsgedächtnisses* (Vol. 10, pp. 37-58). Göttingen, Germany: Hogrefe.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*(1), 29-51. doi: 10.1037/0033-2909.95.1.29

Fox, J.-P., Entink, R. K., & van der Linden, W. (2007). Modeling of Responses and Response Times with the Package cirt. *Journal of Statistical Software, 20*(7), 14. doi: 10.18637/jss.v020.i07

Fox, M. C., & Mitchum, A. L. (2013). A knowledge-based theory of rising scores on “culture-free” tests. *Journal of Experimental Psychology: General, 142*(3), 979–1000. doi: 10.1037/a0030155

Freeman, J. (1979). *Gifted children: Their identification and development in social context*. Lancaster, UK: MTP.

Gagné, F. (1993). Constructs and models pertaining to exceptional human abilities. In K. A. Heller & F. J. Mönks (Eds.), *International handbook of research and development of giftedness and talent* (pp. 69-87). Oxford: Pergamon.

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence, 39*(2–3), 108–119. doi: 10.1016/j.intell.2011.02.001

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not Always Better: The Relation between Item Response and Item Response Time in Raven’s Matrices. *Journal of Intelligence, 3*(1), 21. doi: 10.3390/jintelligence3010021

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*(1), 13-23. doi: 10.1016/S0160-2896(97)90011-8

Graue, E. B. (1985). *Is your child gifted? A handbook for parents of gifted preschoolers*. San Diego: Oak Tree.

Green, K. E., & Kluever, R. C. (1992). Components of Item Difficulty of Raven's Matrices. *Journal of General Psychology, 119*(2), 189.

Guilford, J. P. (1957). Creative abilities in the arts. *Psychological Review, 64*(2), 110-118. doi: 10.1037/h0048280

Harrison, T. L., Shipstead, Z., & Engle, R. W. (2014). Why is working memory capacity related to matrix reasoning tasks? *Memory & Cognition, 43*(3), 389-396. doi: 10.3758/s13421-014-0473-3

Hasselhorn, M., & Schuchardt, K. (2006). Lernstörungen: Eine kritische Skizze zur Epidemiologie. *Kindheit und Entwicklung, 15*(4), 208-215. doi: 10.1026/0942-5403.15.4.208

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2011). A novel method for analyzing sequential eye movements reveals strategic influence on Raven's Advanced Progressive Matrices. *Journal of vision, 11*(10), 10-10. doi: 10.1167/11.10.10

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence, 48*(0), 1-14. doi: 10.1016/j.intell.2014.10.005

Heller, K. A., Perleth, C., & Hany, E. A. (1994). Hochbegabung - ein lange Zeit vernachlässigtes Forschungsthema. *Forschung an der Ludwig-Maximilians-Universität, 3*(1), 18-22.

Helmsen, J., Lehmkuhl, G., & Petermann, F. (2009). Kinderpsychiatrie und Klinische Kinderpsychologie im Dialog. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie, 57*(4), 285-296. doi: 10.1024/1661-4747.57.4.285

Hitch, G. J., & Baddeley, A. D. (1976). Verbal reasoning and working memory. *Quarterly Journal of Experimental Psychology, 28*(4), 603-621. doi: 10.1080/14640747608400587

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*(2), 220.

Hunt, E. (1974). Quote the Raven? Nevermore. In L. W. Gregg (Ed.), *Knowledge and cognition* (pp. 129-157). Oxford, UK: Lawrence Erlbaum.

Jackson, N. E., Donaldson, G. W., & Cleland, L. N. (1988). The structure of precocious reading ability. *Journal of Educational Psychology, 80*(2), 234-243. doi: 10.1037/0022-0663.80.2.234

Jarosz, A. F., & Wiley, J. (2012). Why does working memory capacity predict RAPM performance? A possible role of distraction. *Intelligence, 40*(5), 427–438. doi: 10.1016/j.intell.2012.06.001

Jensen, A. R. (1993). Why is reaction time correlated with psychometric g? *Current Directions in Psychological Science, 2*(2), 53-56. doi: 10.2307/20182199

Johnson-Laird, P. N. (2004). The history of mental models. In K. Manktelow & M. C. Chung (Eds.), *Psychology of Reasoning. Theoretical and Historical Perspectives* (pp. 179-212). Hove, UK: Psychology Press.

Johnson-Laird, P. N. (2005). Mental Models and Thought. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 185-208). New York: Cambridge University Press.

Jung, E., Molfese, V. J., Beswick, J., Jacobi-Vessels, J., & Molnar, A. (2009). Growth of Cognitive Skills in Preschoolers: Impact of Sleep Habits and Learning-Related Behaviors. *Early Education & Development, 20*(4), 713-731. doi: 10.1080/10409280802206890

Kandel, E. R. (2001). The Molecular Biology of Memory Storage: A Dialogue Between Genes and Synapses. *Science, 294*(5544), 1030-1038. doi: 10.1126/science.1067020

Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory as variation in executive attention and control. In A. R. A.

Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21-48). Oxford, England: Oxford University Press.

Kaufman, A. S., Kaufman, N. L., Melchers, P., & Preuß, U. (2009). *K-ABC: Kaufman – Assessment Battery for Children* (8th ed.). Frankfurt: Pearson.

Kim, K. H. (2005). Can Only Intelligent People Be Creative? A Meta-Analysis. *Journal of Advanced Academics, 16*(2-3), 57-66. doi: doi:10.4219/jsge-2005-473

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., . . . Pandey, A. (2014). A draft map of the human proteome. *Nature, 509*(7502), 575-581. doi: 10.1038/nature13302

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-25. Retrieved from http://CRAN.R-project.org/package=lmerTest

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*(4), 389-433. doi: 10.1016/s0160-2896(05)80012-1

Lehwald, G. (1991). Curiosity and exploratory behaviour in ability development. *European Journal of High Ability, 1*(2), 204-210. doi: 10.1080/0937445910010212

Lewis, M., & Michalson, L. (1985). The gifted infant. In J. Freeman (Ed.), *The Psychology of gifted children : perspectives on development and education* (pp. 35-57). Chichester: John Wiley & Sons.

Linden, D. E. J. (2007). The Working Memory Networks of the Human Brain. *The Neuroscientist, 13*(3), 257-267. doi: 10.1177/1073858406298480

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

Marland, S. P. (1971). *Education of the gifted and talented volume 1: Report to the congress of the United States by the US commissioner of education.* Washington, DC: U.S. Department of Health, Education & Welfare.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7*(2), 107–127. doi: 10.1016/0160-2896(83)90023-5

Martínez, K., Burgaleta, M., Román, F. J., Escorial, S., Shih, P. C., Quiroga, M. Á., & Colom, R. (2011). Can fluid intelligence be reduced to 'simple' short-term storage? *Intelligence, 39*(6), 473–480. doi: 10.1016/j.intell.2011.09.001

Marx, H. (1992). *Vorhersage von Lese-Rechtschreibschwierigkeiten in Theorie und Anwendung*. Habilitationsschrift. Universität Bielefeld.

Matzen, L. E., Benz, Z. O., Dixon, K. R., Posey, J., Kroger, J. K., & Speed, A. E. (2010). Recreating Raven's: software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods, 42*(2), 525-541. doi: 10.3758/brm.42.2.525

McCall, R. B., & Carriger, M. S. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development, 64*(1), 57-79. doi: 10.1111/j.1467-8624.1993.tb02895.x

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. doi: 10.1016/j.intell.2008.08.004

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220–232. doi: 10.1037/h0048850

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. doi: 10.1037/a0019182

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General, 130*(4), 621–640. doi: 10.1037/0096-3445.130.4.621

Moscovitch, M., Rosenbaum, R. S., Gilboa, A., Addis, D. R., Westmacott, R., Grady, C., . . . Nadel, L. (2005). Functional neuroanatomy of remote episodic, semantic and spatial memory: a unified account based on multiple trace theory. *Journal of Anatomy, 207*(1), 35-66. doi: 10.1111/j.1469-7580.2005.00421.x

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12*(2), 252–284. doi: 10.1016/0010-0285(80)90011-0

Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77-101. doi: 10.1037/0003-066x.51.2.77

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. doi: 10.1016/0191-8869(90)90007-E

Nusbaum, E. C., & Silvia, P. J. (2011). Are intelligence and creativity really so different? Fluid intelligence, executive processes, and strategy use in divergent thinking. *Intelligence, 39*(1), 36–45. doi: 10.1016/j.intell.2010.11.002

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working Memory and Intelligence - Their Correlation and Their Relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 61–65. doi: 10.1037/0033-2909.131.1.61

Oberauer, K., Süß, H.-M., Wilhelm, O., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), Variation in Working Memory. New York: Oxford University Press.

Oberauer, K., Weidenfeld, A., & Hörnig, R. (2006). Working memory capacity and the construction of spatial mental models in comprehension and deductive reasoning. *The Quarterly Journal of Experimental Psychology, 59*(2), 426–447. doi: 10.1080/17470210500151717

Perleth, C. (2000). Neue Tendenzen und Ergebnisse in der Begabungs- und Intelligenzdiagnostik. In H. Joswig (Ed.), *Begabung erkennen - Begabte fördern* (pp. 35-64). Rostock: Uni Rostock.

Perleth, C., Schatz, T., & Mönks, F. J. (2000). Early identification of high ability. In K. A. Heller, F. J. Mönks, R. J. Sternberg, & R. F. Subotnik (Eds.), *International Handbook of Giftedness and Talent* (2nd ed., pp. 297-316). Oxford, UK: Pergamon.

Pollock, J. I. (1992). Predictors and long-term associations of reported sleeping difficulties in infancy. *Journal of Reproductive and Infant Psychology, 10*(3), 151-168. doi: 10.1080/02646839208403947

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30*(1), 41-70. doi: 10.1016/s0160-2896(01)00067-8

Primi, R. (2002). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30*(1), 41–70. doi: 10.1016/s0160-2896(01)00067-8

R Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93-134. doi: 10.1037/0033-295X.88.2.93

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Pearson.

Revelle, W. (2015). psych: Procedures for Personality and Psychological Research (Version 1.5.8). Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych

Roberts, R. D., & Stankov, L. (1999). Individual differences in speed of mental processing and human cognitive abilities: Toward a taxonomic model. *Learning and Individual Differences, 11*(1), 1-120. doi: 10.1016/S1041-6080(00)80007-2

Robinson, N. M., Dale, P. S., & Landesman, S. (1990). Validity of Stanford-Binet IV with linguistically precocious toddlers. *Intelligence, 14*(2), 173-186. doi: 10.1016/0160-2896(90)90003-c

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2005). The structure of infant cognition at 1 year. *Intelligence, 33*(3), 231-250. doi: 10.1016/j.intell.2004.11.002

Rose, S. A., Feldman, J. F., Jankowski, J. J., & Van Rossem, R. (2012). Information processing from infancy to 11 years: Continuities and prediction of IQ. *Intelligence, 40*(5), 445-456. doi: 10.1016/j.intell.2012.05.007

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General, 126*(3), 211-227. doi: 10.1037/0096-3445.126.3.211

Rosenzweig, M. R., Breedlove, S. M., & Watson, N. V. (2005). *Biological Psychology: An Introduction to Behavioral and Cognitive Neuroscience* (4th ed.). Sunderland, MA: Sinauer.

Rost, D. H. (1993). Das Marburger Hochbegabtenprojekt. In D. H. Rost (Ed.), *Lebensumweltanalyse hochbegabter Kinder: Das Marburger Hochbegabtenprojekt* (pp. 1-33). Göttingen: Hogrefe.

Rost, D. H. (2009a). Grundlagen, Fragestellungen, Methode. In D. H. Rost (Ed.), *Hochbegabte und hochleistende Jugendliche* (2nd ed., pp. 1-91). Münster: Waxmann.

Rost, D. H. (2009b). *Intelligenz: Fakten und Mythen*. Weinheim: Beltz.

Rubin, R. A., & Balow, B. (1979). Measures of infant development and socioeconomic status as predictors of later intelligence and school achievement. *Developmental Psychology, 15*(2), 225-227. doi: 10.1037/0012-1649.15.2.225

Rudolf, H. (1980). Die Entwicklung der Graphomotorik als psychomotorischer Prozess. *Diagnostica, 26*(4), 354-360.

Salthouse, T., & Pink, J. (2008). Why is working memory related to fluid intelligence? *Psychonomic Bulletin & Review, 15*(2), 364-371. doi: 10.3758/pbr.15.2.364

Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Sattler.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the Relation between Time on Task and Ability in Complex Problem Solving. *Intelligence, 48*, 37-50. doi: 10.1016/j.intell.2014.10.003

Schneider, W., Bullock, M., & Sodian, B. (1998). Die Entwicklung des Denkens und der Intelligenzunterschiede zwischen Kindern. In F. E. Weinert (Ed.), *Entwicklung im Kindesalter* (pp. 53 - 74). Weinheim: Psychologie Verlags Union.

Schweizer, K., & Koch, W. (2001). A revision of Cattell's Investment Theory: Cognitive properties influencing learning. *Learning and Individual Differences, 13*(1), 57-82. doi: 10.1016/S1041-6080(02)00062-6

Shapiro, B. K., Palmer, F. B., Antell, S. E., Bilker, S., Ross, A., & Capute, A. J. (1989). Giftedness: Can it be predicted from infancy? . *Clinical Pediatrics, 28*(5), 205-209. doi: 10.1177/000992288902800501

Shiffrin, R. M., & Atkinson, R. C. (1969). Storage and retrieval processes in long-term memory. *Psychological Review, 76*(2), 179-193. doi: 10.1037/h0027277

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., . . . Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts, 2*(2), 68–85. doi: 10.1037/1931-3896.2.2.68

Slater, A. (1997). Can measures of infant habituation predict later intellectual ability? *Archives of Disease in Childhood, 77*(6), 474-476. doi: 10.1136/adc.77.6.474

Snow, R. E. (1980). Aptitude Processes. In R. E. Snow, P.-A. Federico, & W. E. Montague (Eds.), *Aptitude, Learning, and Instruction. Volume 1: Cognitive Process Analyses of Aptitude* (pp. 27-63). Hillsdale, NJ: Erlbaum.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology, 15*(2), 201–292. doi: 10.2307/1412107

Spearman, C. (1923). *The Nature of Intelligence and the Principles of Cognition*. London, UK: Macmillan.

Spearman, C. (1938). Measurement of intelligence. *Scientia*, *64*, 75-82.

Stamm, M. (2004). Lernentwicklungen von Frühlesern und Frührechnerinnen. *Zeitschrift für Erziehungswissenschaft, 7*(3), 395-415. doi: 10.1007/s11618-004-0041-x

Stapf, A. (2010). *Hochbegabte Kinder. Persönlichkeit, Entwicklung, Förderung* (5th ed.). München: C.H.Beck.

Stapf, A., & Stapf, K. H. (1988). Kindliche Hochbegabung in entwicklungs-psychologischer Sicht. *Zeitschrift für Psychologie in Erziehung und Unterricht, 35*, 1-17.

Sternberg, R. J. (1985). *Beyond IQ : a triarchic theory of human intelligence*. Cambridge: Cambridge University Press.

Sternberg, R. J. (1986). Haste makes waste versus a stitch in time? A reply to Vernon, Nador, and Kantor. *Intelligence*, *10*(3), 265-270. doi: 10.1016/0160-2896(86)90020-6

Stöger, H., Schirner, S., & Ziegler, A. (2008). Ist die Identifikation Begabter schon im Vorschulalter möglich? Ein Literaturüberblick. *Diskurs Kindheits- und Jugendforschung, 3*(1), 7-24.

Styles, I., Raven, M., & Raven, J. C. (1998). *Raven's Progressive Matrices SPM Plus Sets A-E*. San Antonio, TX: Harcourt.

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability - and a little bit more. *Intelligence, 30*(3), 261-288. doi: 10.1016/s0160-2896(01)00100-3

Terassier, J. C. (1985). Dyssynchrony: Uneven Development. In J. Freeman (Ed.), *The psychology of gifted children: Perspectives on development and education* (pp. 265-274). Chistester: Wiley.

Terman, L. M. (1925). *Genetic Studies of Genius*. Stanford University Press.

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research, 76*, 31-42. doi: 10.1016/j.visres.2012.10.012

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *The Stanford-Binet intelligence scale, fourth edition: Guide for administering and scoring* (4th ed.). Chicago, IL: Riverside.

Thurstone, L. L. (1938). *Primary Mental Abilities*. Chicago, IL: The University of Chicago Press.

Topolinski, S., & Reber, R. (2010). Gaining Insight Into the “Aha” Experience. *Current Directions in Psychological Science, 19*(6), 402-405. doi: 10.1177/0963721410388803

Turley-Ames, K. J., & Whitfield, M. M. (2003). Strategy training and working memory task performance. *Journal of Memory and Language, 49*(4), 446-468. doi: 10.1016/S0749-596X(03)00095-0

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Working memory capacity and retrieval from long-term memory: the role of controlled search. *Memory & Cognition, 41*(2), 242-254. doi: 10.3758/s13421-012-0261-x

Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between Operation Span and Raven. *Intelligence, 33*(1), 67–81. doi: 10.1016/j.intell.2004.08.003

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A Dynamical Model of General Intelligence: The Positive Manifold of Intelligence by Mutualism. *Psychological Review, 113*(4), 842-861. doi: 10.1037/0033-295X.113.4.842

van der Maas, H. L. J., Kan, K.-J., & Borsboom, D. (2014). Intelligence Is What the Intelligence Test Measures. Seriously. *Journal of Intelligence, 2*(1), 12-15. doi: 10.3390/jintelligence2010012

Vandierendonck, A. (2016). A Working Memory System With Distributed Executive Control. *Perspectives on Psychological Science, 11*(1), 74-100. doi: 10.1177/174591615596790

Verguts, T., De Boeck, P., & Maris, E. (1999). Generation speed in Raven's progressive matrices test. *Intelligence, 27*(4), 329-345. doi: 10.1016/s0160-2896(99)00023-9

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

Vock, M., & Holling, H. (2008). The measurement of visuo–spatial and verbal–numerical working memory: Development of IRT-based scales. *Intelligence, 36*(2), 161–182. doi: 10.1016/j.intell.2007.02.004

Wechsler, D. (2009). Manual. In M. von Aster, A. C. Neubauer, & R. Horn (Eds.), *Wechsler Intelligenztest für Erwachsene* (2 ed.). Frankfurt am Main, Germany: Pearson Assessment.

Wechsler, D., & Petermann, F. (2009). *Wechsler Preschool and Primary Scale of Intelligence III - Deutsche Version*. Frankfurt am Main, Germany: Pearson.

Weiner, B. (1985). An Attributional Theory of Achievement Motivation and Emotion. *Psychological Review, 92*(4), 548-573. doi: 10.1037/0033-295X.92.4.548

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

Winner, E. (2004). *Hochbegabte: Mythen und Realitäten von außergewöhnlichen Kindern*. Stuttgart: Klett-Cotta.

Appendix A: Supplementary Material

Supplementary data files for the experiments in Part 2 are available for download at http://dx.doi.org/10.1016/j.intell.2014.10.004

Supplementary data files and R-scripts for the experiments in Part 3 are available at request from the author (loesche@dipf.de).

Appendix B: Documents

Erklärung zu bisherigen Promotionsverfahren

Ich erkläre hiermit, dass ich mich bisher keiner Doktorprüfung im Mathematisch-Naturwissenschaftlichen Bereich unterzogen habe.

Frankfurt am Main, den 12. April 2016________________________________

Dipl.-Psych. Patrick Lösche

Erklärung zur Promotionsordnung

Ich erkläre hiermit, dass mir die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fachbereiche der Goethe Universität Frankfurt am Main bekannt ist.

Frankfurt am Main, den 12. April 2016________________________________

Dipl.-Psych. Patrick Lösche

Eidesstattliche Versicherung

Ich erkläre hiermit, dass ich die vorgelegte Dissertation mit dem Titel „*The Role of Rule Knowledge in Inductive Reasoning*“ selbständig angefertigt und mich nicht anderer Hilfsmittel als der in ihr angegebenen bedient habe, insbesondere, dass alle Entlehnungen aus anderen Schriften mit Angabe der betreffenden Schrift gekennzeichnet sind.

Ich versichere, die Grundsätze der guten wissenschaftlichen Praxis beachtet, und nicht die Hilfe einer kommerziellen Promotionsvermittlung in Anspruch genommen zu haben.

Frankfurt am Main, den 12. April 2016________________________________

Dipl.-Psych. Patrick Lösche

Erfüllung der Kriterien für publikationsbasierte Dissertation

Seite 1

Ich erkläre nachfolgend, die Kriterien für kumulative Dissertationen (Punkte 1-7) zu erfüllen, die für den Fachbereich Psychologie und Sportwissenschaften der Goethe Universität Frankfurt gültig sind (nach Beschluss im Fachbereichsrat ab 11.6.2015).

(1) Die kumulative Dissertation soll in der Regel 3 Schriften umfassen, die aus den letzten 5 Jahren stammen sollen.

Erklärung: Die Dissertation umfasst drei Schriften aus den Jahren 2014 bis 2016.

(2) Die Schriften sollen im Wesentlichen einem zusammenhängenden Forschungsprogramm entstammen. Die jeweils verfolgten Forschungsfragen sollen sich sinnvoll zueinander in Beziehung setzen lassen.

Erklärung: Die Schriften entstammen im Wesentlichen einem zusammenhängenden Forschungsprogramm. Die Schriften stehen alle in Verbindung zu der Fragestellung über die grundlegenden kognitiven Prozesse der Intelligenz.

(3) Der Kandidat oder die Kandidatin soll bei 2 Publikationen Erstautor/Erstautorin sein, bei einer weiteren Publikation kann er/sie Koautor/Koautorin sein. Eine geteilte Erstautorenschaft wird für jeden der Erstautoren anteilig gewichtet (bei 2 Erstautoren eine 1/2 Erstautorenschaft, bei 3 eine 1/3 Erstautorenschaft usw.).

Erklärung: Der Verfasser ist bei zwei der drei Schriften Erstautor und Koautor bei einer weiteren Schrift. Es gibt keine geteilten Erstautorenschaften. Damit ist das Kriterium von mindestens zwei Erstautorenschaften erfüllt.

Erfüllung der Kriterien für publikationsbasierte Dissertation

Seite 2

(4) Die drei Schriften sollen zur Veröffentlichung zumindest eingereicht sein. Der aktuelle Status ist detailliert darzulegen (Publikationsorgan und Status wie eingereicht, in revision, conditional accept usw.).

(5) Mindestens 2 der 3 Schriften müssen in guten oder sehr guten, in der Regel englischsprachigen, Zeitschriften mit Peer-Review eingereicht sein.

(6) Eine der 3 Schriften kann als Publikation in einem einschlägigen Lehrbuch, Enzyklopädieband oder einem anderen für das jeweilige Fach bedeutsamen

Erklärung: Eine der Schriften wurde 2014 als Kapitel in dem Sammelwerk „*Handbuch Talententwicklung*“ veröffentlicht. Eine weitere Schrift wurde 2015 in der internationalen Fachzeitschrift „*Intelligence*“ veröffentlicht. Eine dritte Schrift wurde am 3.3.2016 bei der internationalen Fachzeitschrift „*Journal of Experimental Psychology: Learning Memory and Cognition*“ eingereicht und erwartet eine Begutachtung.

(7) Die als Dissertation vorgelegte Abhandlung soll über die zusammengestellten Publikationen hinaus einen zusätzlichen Text enthalten, in welchem eine kritische Einordnung der eigenen Publikationen aus einer übergeordneten Perspektive heraus vorgenommen wird. Dieser Text sollte einen Umfang von ca. 30 Seiten haben. Es sollen die Fragestellungen theoretisch entwickelt werden, die empirischen Arbeiten und ihre Ergebnisse so dargestellt werden, dass sie auch ohne Lesen der Einzelarbeiten nachvollziehbar sind und es soll eine Gesamtdiskussion enthalten, die die Fragestellungen beantwortet und den Erkenntnisgewinn der Arbeit herausstellt.

Erklärung: Ein entsprechender Text ist enthalten.

Frankfurt, den 12. April 2016 ______________________________

Dipl.-Psych. Patrick Lösche

Erklärung über die Eigenleistung

Seite 1

Die vorliegende Dissertation mit dem Titel „*The Role of Rule Knowledge in Inductive Reasoning*“ beinhaltet drei Manuskripte, die zum Zeitpunkt der Eröffnung des Promotionsverfahrens in internationalen Fachzeitschriften und Sammelwerken veröffentlicht oder zur Veröffentlichung eingereicht wurden. Alle drei Manuskripte (nachfolgend aufgeführt) entstanden unter Mitwirkung des Verfassers, davon in zwei Fällen als Erstautor.

1) Beißert, H., Hasselhorn, M., & Lösche, P. (2014). Möglichkeiten und Grenzen der Frühprognose von Hochbegabung. In M. Stamm (Ed.), *Handbuch Talententwicklung* (pp. 415-425). Bern, Switzerland: Hans Huber.

Das o.g. Manuskript entstand in Zusammenarbeit mit Hanna Beißert und Marcus Hasselhorn. Erstautorin Hanna Beißert hat einen ersten Entwurf gestaltet und war hauptsächlich für das zweite und dritte Unterkapitel verantwortlich. Marcus Hasselhorn schrieb die Einleitung und war am vierten Unterkapitel beteiligt. Der Verfasser hat als Koautor das letzte Unterkapitel geschrieben und war am vierten Unterkapitel beteiligt. Alle Autoren haben das gesamte Manuskript diskutiert und überarbeitet.

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

Das o.g. Manuskript entstand in Zusammenarbeit mit Marcus Hasselhorn und Jennifer Wiley. Der Verfasser war als Erstautor federführend für das Abfassen des Manuskripts verantwortlich, dessen Konzeption und Argumentationskette er selbstständig erarbeitet hat. Er entwickelte die Fragestellung, sammelte die Daten, und führte die Datenauswertung eigenständig durch. Die Ergebnisse wurden mit allen Koautoren ausführlich besprochen. Die Koautoren überarbeiteten das abgefasste Manuskript inhaltlich und sprachlich.

Erklärung über die Eigenleistung

Seite 2

3) Loesche, P. & Hasselhorn, M. (submitted). The Effects of Rule Knowledge on Eye Movements and Response Time in Matrix Reasoning.

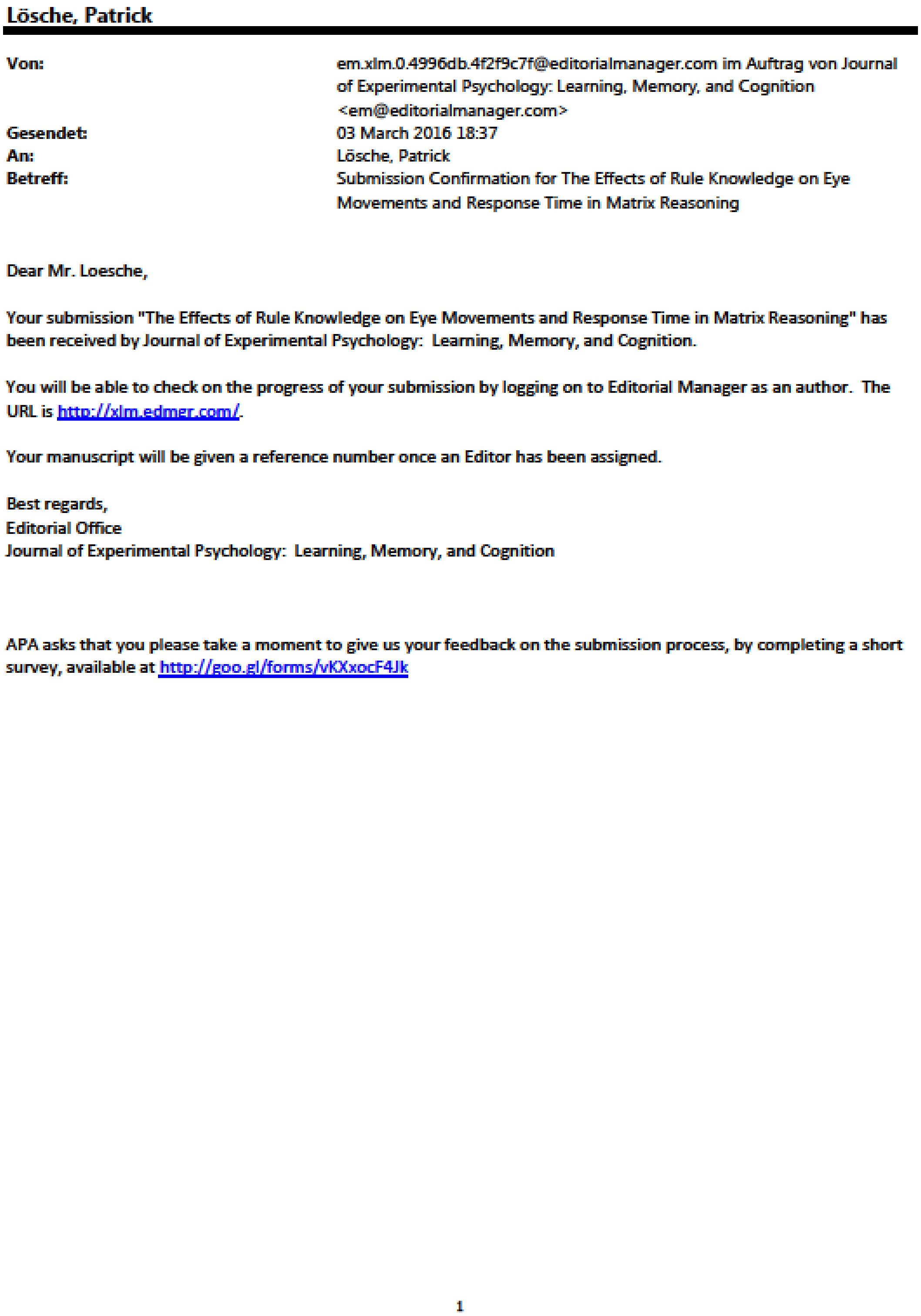
Das o.g. Manuskript entstand in Zusammenarbeit mit Marcus Hasselhorn. Es wurde am 3.3.2016 bei dem *Journal of Experimental Psychology: Learning, Memory, and Cognition* zur Veröffentlichung eingereicht und erwartet eine Begutachtung. Der Verfasser war als Erstautor federführend für das Abfassen des Manuskripts verantwortlich, dessen Konzeption und Argumentationskette er selbstständig erarbeitet hat. Er entwickelte die Fragestellung, sammelte die Daten, und führte die Datenauswertung eigenständig durch. Die Ergebnisse wurden mit dem Koautor ausführlich besprochen. Der Koautor überarbeitete das abgefasste Manuskript inhaltlich und sprachlich.

Frankfurt am Main, den 12. April 2016

Verfasser der Dissertation: ____________________________
Dipl.-Psych. Patrick Lösche

Betreuer der Dissertation: ____________________________
Prof. Dr. Marcus Hasselhorn

Bestätigung der Einreichung



Lösche, Patrick

Von: em.xlm.0.4996db.4f2f9c7f@editorialmanager.com im Auftrag von Journal of Experimental Psychology: Learning, Memory, and Cognition <em@editorialmanager.com>
Gesendet: 03 March 2016 18:37
An: Lösche, Patrick
Betreff: Submission Confirmation for The Effects of Rule Knowledge on Eye Movements and Response Time in Matrix Reasoning

Dear Mr. Loesche,

Your submission "The Effects of Rule Knowledge on Eye Movements and Response Time in Matrix Reasoning" has been received by Journal of Experimental Psychology: Learning, Memory, and Cognition.

You will be able to check on the progress of your submission by logging on to Editorial Manager as an author. The URL is http://xlm.edmer.com/.

Your manuscript will be given a reference number once an Editor has been assigned.

Best regards,
Editorial Office
Journal of Experimental Psychology: Learning, Memory, and Cognition

APA asks that you please take a moment to give us your feedback on the submission process, by completing a short survey, available at http://goo.gl/forms/vKXxocF4Jk

LEBENSLAUF

SEITE 1

Patrick Lösche, Dipl.-Psych.
Geboren am 3.6.1983 in Lübbecke

ADRESSE:
Mittlerer Hasenpfad 38
60598 Frankfurt

☎ 069-24708-240
✉ loesche@dipf.de

BILDUNG:

2002: Abitur am Söderblom Gymnasium, Espelkamp

2003-2010: Georg-August-Universität, Göttingen, Diplom in Psychologie.
Diplomarbeit: *Über den Einfluss von Verantwortlichkeit auf die Lösungswahrscheinlichkeit im Hidden Profile* unter der Betreuung von Prof. Dr. Stefan Schulz-Hardt

2014: Visiting Scholar für drei Monate an der University of Illinois, Chicago, USA.
Unter der Betreuung von Prof. Jennifer Wiley, PhD.

BERUF:

Seit 2010: German Institute for International Educational Research DIPF, Frankfurt

- Wissenschaftlicher Mitarbeiter und Promotionsstudent in der Abteilung für Bildung und Entwicklung unter der Betreuung von Prof. Dr. Marcus Hasselhorn
- Dissertation Arbeitstitel: “What one intelligence test measures. The distinct roles of working memory and rule induction in the Raven progressive matrices”

VERÖFFENTLICHUNGEN:

- Loesche, P., Wiley, J., & Hasselhorn, M. (2015). How knowing the rules affects solving the Raven Advanced Progressive Matrices Test. *Intelligence*, 48, 58-75. doi: 10.1016/j.intell.2014.10.004
- Beißert, H., Hasselhorn, M., & Lösche, P. (2014). Möglichkeiten und Grenzen der Frühprognose von Hochbegabung. In M. Stamm (Ed.), *Handbuch Talententwicklung* (pp. 415-425). Bern, Switzerland: Hans Huber.

LEBENSLAUF

SEITE 2

Patrick Lösche, Dipl.-Psych**.**
Geboren am 3.6.1983 in Lübbecke

VORTRÄGE:

- Loesche, P. & Hasselhorn, M. (2015). *Augenbewegungen beim Lösen figuraler Matrizenaufgaben bei bekannten und unbekannten Lösungsregeln*. Präsentiert bei der 13. Arbeitstagung der Fachgruppe DPPD, Mainz, Germany.
- Loesche, P. & Hasselhorn, M. (2015). *How knowing the rules affects solving the Raven Progressive Matrices test*. Präsentiert bei der 57. Tagung experimentell arbeitender Psychologen, Hildesheim, Germany.
- Loesche, P. & Hasselhorn, M. (2014). *The Role of Rule Induction and Working Memory in Matrix Reasoning*. Präsentiert bei dem 49. Kongress der Deutschen Gesellschaft für Psychologie, Bochum, Germany.
- Loesche, P. & Hasselhorn, M. (2011). *What do matrices tests measure?* Präsentiert bei der 19th Biennial World Conference of the World Council for Gifted and Talented Children (WCGTC), Prague, Czech.

LEHRE:

WS 2011/12: “Psychologische Aspekte der Hochbegabung” (Seminar), Goethe Universität, Frankfurt, Germany.