Tracking the Decision Making Process in Multiple-Choice Assessment:

Evidence from Eye Movements

Marlit Annalena Lindner[a]*, Alexander Eitel[b],

Gun-Brit Thoma[a], Inger-Marie Dalehefte[a], Jan Marten Ihme[a], Olaf Köller[a]

PLEASE CITE AS:

Lindner, M. A., Eitel, A., Thoma, G.-B., Dalehefte, I.-M., Ihme, J. M. & Köller, O. (2014). Tracking the decision making process in multiple-choice assessment: Evidence from eye movements. *Applied Cognitive Psychology*, *28*(5), 738–752. doi:10.1002/acp.3060

See also: http://onlinelibrary.wiley.com/doi/10.1002/acp.3060/abstract

*Correspondence concerning this article should be addressed to Marlit Annalena Lindner, Leibniz-Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany. Tel: +49 431 880 4410; Fax: +49 431 880 2629;

Email: mlindner@ipn.uni-kiel.de

Abstract

This study investigated students' decision making processes in a knowledge-assessing multiple-choice (MC) test using eye tracking methodology. More precisely, the gaze bias effect (more attention to more preferred options) and its relation to domain knowledge were the focus of the study. Eye movements of students with high (HPK) and low (LPK) prior domain knowledge were recorded while they solved 21 MC items. Afterwards, students rated every answer option according to their subjective preference. As expected, both HPK and LPK students showed a gaze bias towards subjectively preferred answer options, whereby HPK students spent more time on objectively correct answers. Furthermore, a fine-grained time course analysis showed similar patterns of attention distribution over time for both HPK and LPK students, when focusing on subjective preference levels. Thus, these data offer a new perspective on knowledge-related MC item-solving and provide evidence for the generalizability of the gaze bias effect across decision tasks.

Keywords: multiple-choice questions; decision making; eye tracking; gaze bias effect; cognitive diagnostic assessment

Tracking the Decision Making Process in Multiple-Choice Assessment: Evidence from Eye Movements

Multiple-choice (MC) questions are acknowledged as having remarkably positive characteristics in educational assessment due to their ease of use in practical application, especially in terms of standardization and item scoring (Haladyna, 2004). Therefore, MC items are frequently used to assess knowledge in everyday educational settings as well as in prominent large-scale studies such as PISA - *Programme for International Student Assessment* (OECD, 2013). Due to the political power of such educational studies and due to the use of high-stakes tests as an admission restriction measure in many educational systems, developing high quality assessments is crucial and needs to be accompanied by solid research. Accordingly, constructional aspects of item writing and psychometrical issues have received much attention in the last decades, while only few studies have applied a cognitive perspective to students' demands and processing when they solve MC items or related assessments. Such knowledge, however, could be particularly useful in future research on item characteristics and their interaction with students' characteristics (Embretson, 1999; Haladyna, Downing, & Rodriguez, 2002; Leighton, 2004) to possibly increase test fairness and the validity of assessments. This is a central goal in the field of *cognitive diagnostic assessment* (CDA).

By using insights and methodology (i.e., eye tracking) from cognitive psychology to explore students' processing of MC items in educational settings, the present study was conducted to possibly support future efforts in CDA. In particular, against the backdrop of theory and research on knowledge-related cognitive processing (e.g., Canham & Hegarty, 2010; Sweller, Van Merriënboer, & Paas, 1998) and decision making (e.g., Glaholt, Wu, & Reingold, 2009; Shimojo, Simion, Shimojo, & Scheier, 2003), in the present study, we derived three hypotheses about students' processing and solving of MC items. To test the hypotheses, we conducted fine-grained quantitative analyses of high prior knowledge (HPK) and low prior knowledge (LPK) students' eye movements during the item-solving process. The results may contribute to a better understanding of how students with different levels of domain knowledge process MC test information. Furthermore, they provide tentative evidence for the potential of using eye tracking data to assess students' domain knowledge levels and preferences for answer options, indicating that eye tracking data could be used as diagnostic information in future educational practice.

1 Theoretical Background

1.1 The Potential of Eye Tracking in Cognitive Diagnostic Assessment

Cognitive diagnostic assessment means combining various methods and theoretical approaches (e.g., from cognitive psychology) to develop and improve tests. For instance, students' processing data (e.g., verbal protocols) can be consulted to examine and account for the construct and internal validity of tests or to allow for a more valid diagnosis of students' abilities and needs in future instruction (Embretson & Gorin, 2001; Leighton, 2004; Messick, 1989; Nichols, 1994). Another goal of CDA is to develop and test cognitive models to explain the item-solving process and, thus, to provide solid groundwork for more theory-driven test construction and, consequently, higher test quality in the future (e.g., Leighton & Gierl, 2007). Thus, taken together, the main intention of CDA is to learn more about the cognitive requirements of test items in educational assessment by making use of statistical models, cognitive theories, and methods to gain insights into the item-solving process (Healy, 2005; Leighton & Gierl, 2007; Nichols, Chipman, & Brennan, 1995; Pellegrino, Chudowsky, & Glaser, 2001).

A sophisticated method to obtain such processing data in testing situations is the use of eye tracking technology (for an introduction, see Duchowski, 2007; Holmqvist, Nyström et al., 2011). This method is perfectly suited (not only) for the context of MC assessment because it combines several advantages that allow high quality processing data to be acquired: First, compared to traditional process tracing methods (e.g., think aloud protocols), eye tracking allows students' attention distribution to be recorded while they solve a task without placing any extra load on participants' working memory (e.g., Lohse & Johnson, 1996; Russo, 1978). Furthermore, eye movement recordings are objective and provide high-frequency data of both temporal and spatial information of eye movement behavior while having the potential to reveal even unconscious cognitive events not accessible from self-reports or external observation (e.g., the *gaze bias effect*; see below). As a precondition for the interpretation of eye movement patterns, the eye-mind hypothesis (Just & Carpenter, 1980) assumes that attention is focused on the point of fixation so that eye movements reflect the spatiotemporal encoding of visual information, thus providing a valid indirect measure of attention distribution and cognitive processing. Even though shortcomings of the eye-mind hypothesis have been discussed (Engbert & Kliegl, 2003; Hyönä, 2010; Posner, 1980; Posner, Snyder, & Davidson 1980; Wright & Ward, 2008), under conditions of natural viewing with an engaging task (like solving an MC test), attention and gaze are closely coupled (cf., Holmqvist, Nyström et al., 2011). This is also supported by neurophysiological data (e.g., Kustov & Robinson, 1996).

Accordingly, when solving MC items, eye movements can provide a valuable instrument to track the cognitive processes of information acquisition and decision making behavior, because solving a standard MC test can be considered to be a multi-alternative decision making situation: Students have to choose one best answer option from a set of available options, which consist of the correct answer and several incorrect distractors. Due to the distinct locations of answer options and the item stem in standard MC items (e.g., Figure 2), eye movements can reflect this choice process, which refers to a question or a statement or a problem that is presented in the so-called item stem (Haladyna, 2004). As MC item-solving always requires decision making, in the *first* theoretical section, we will refer to selected research on eye movements in the area of decision making to derive specific hypotheses on how students reach a decision in an MC testing situation.

In general, educational MC tests are specifically constructed to measure students' domain knowledge so that knowledge in the test domain might not only lead to choosing more correct answers in the test, but also to changes in how answer options are processed. Consequently, in a *second* theoretical section, we refer to theories and research on knowledge-related cognitive processing to derive hypotheses about how HPK and LPK students solve MC tests.

1.2 The '*Gaze Bias Effect*' in Decision Making Situations

In the literature on eye movements and decision making it has become a well-established observation that humans have a tendency to shift their attention more towards alternatives they subjectively perceive as being attractive, and thus consider for choice (e.g., Glaholt et al., 2009; Glaholt & Reingold, 2009a, 2009b, 2011; Glaholt, Wu, & Reingold, 2010; Krajbich & Rangel, 2011; Pieters & Warlop, 1999; Schotter, Berry, McKenzie, & Rayner, 2010; Simion & Shimojo, 2006). This effect is known as the '*gaze bias effect*' (or '*gaze cascade effect'*). Shimojo et al. (2003) first described the effect as an increasing shift in attention towards the eventually chosen option in a two-alternative forced choice face-liking task in the final section of the decision process (i.e., two seconds prior to the motor response). This shift of attention towards the chosen option proved to be a stable phenomenon in various decision making situations. Examples are *consumer decision behavior* (Glaholt, Wu, & Reingold, 2010; Pieters & Warlop, 1999; Reutskaja, Nagel, Camerer, & Rangel, 2011), *lineup identification* (Flowe, 2011; Flowe & Cottrell, 2011; Mansour, Lindsay, Brewer, & Munhall, 2009) and *face preference decisions* (Bird, Lauwereyns, & Crawford, 2012; Mitsuda & Glaholt, 2014; Shimojo et al., 2003). One may assume that this attention shift or gaze bias occurs *after* a decision has been made on a conscious level, thus merely reflecting the programming of the motor response. However, simple response-related explanations could be rejected as recent studies show that the bias occurs for a substantial period of time *prior* to the decision announcement (Glaholt & Reingold, 2009a, 2009b, 2011; Simion & Shimojo, 2006). Thus, from a theoretical perspective it is still an open question which specific cognitive processes are reflected by the gaze bias phenomenon.

Nevertheless, relying on gaze bias findings, Glaholt et al. (2009) conducted a study to evaluate the potential of predicting subjective option preferences in decision tasks solely based on gaze parameters. They showed that high positive correlations existed between preference and fixation probability (for whole stimuli as well as single features of stimuli; and even on a person level). Therefore, fixation times during the entire decision making process were longer for the more preferred and especially for the chosen options. Additionally, a gaze likelihood analysis of two seconds prior to the choice announcement allowed for good discrimination between options with distinct preference levels. Bee, Prendinger, Nakasone, André and Ishizuka (2006) even suggested employing the gaze bias effect to automatically detect users' preferences using a newly developed 'AutoSelect' system in the context of human-computer interaction. Their exploratory investigations indeed revealed 81 % correct classification of potential choices in a two-alternative forced choice task, accounting for satisfying stability of the gaze bias effect and high system accuracy. Even in the context of a cognitive problem solving task with a choice component, Ellis, Glaholt and Reingold (2011) provided evidence that eye movements are capable of revealing solution insight prior to students' motor response, and even prior to subjective solution awareness. Taken together, eye movements appear highly related to option preferences and choices in decision making situations.

Since solving MC questions requires decision making (i.e., choosing the correct answer option from a set of options), there is reason to assume that the gaze bias effect, as found in basic decision research, may also occur in the educational context of MC testing. However, in this context, it can be expected that subjective preferences for answer options are closely tied to students' knowledge in the test domain. Hence, subjective preference is likely to be based on existing domain knowledge for HPK students, whereas it is more likely to be based on informed guessing and everyday theories for LPK students. Therefore, subjectively preferred answer options are more likely to be (objectively) correct for HPK than for LPK students. However, even though knowledge levels probably influence subjective preferences, it can be expected that the gaze bias effect occurs as a function of resulting preference (cf., Shimojo et al., 2003), regardless of the *source* of preference.

1.3 Knowledge-Related Cognitive Processing of Multiple-Choice Questions

As described above, domain knowledge is assumed to influence the answer options that students prefer and choose in MC tests. Given that MC items are deliberately constructed to measure the domain knowledge of students, their levels of domain knowledge should also determine whether the chosen answer options are *correct*. But what are the cognitive and attentional processes associated with choosing correct answer options in MC tests? Due to the relative novelty of the application of eye tracking measures in the field of MC, specific hypotheses or theories about how knowledge influences MC item processing are sparse. Thus, in the following, we refer more generally to theories and empirical findings on knowledge-related cognitive processing in order to generate hypotheses on how students solve MC test items.

On a theoretical level, we refer to the conception of human cognitive architecture by Sweller et al. (1998) to explain how knowledge stored in long-term memory affects cognitive processing. According to this conception, persons with high levels of domain knowledge have constructed and automatized schemas about the target domain in long-term memory. Possessing such automatized schemas means that relevant domain knowledge is effectively stored and categorized in long-term memory so that it can be used to solve novel tasks without exceeding working memory resources. In particular, domain knowledge in the form of schemas can foster task performance by optimizing solution-relevant behavior (e.g., Kim & Rehder, 2011). Optimizing solution-relevant behavior can also mean optimizing information processing; that is, by focusing on information that is most relevant to correctly solving the task at hand. Accordingly, empirical eye tracking studies, in which the task was to understand complex weather maps, have shown that students with higher domain knowledge levels (i.e., HPK students) were better able to focus on information that was most relevant for correctly solving the task than students with lower domain knowledge levels (i.e., LPK students). In consequence, HPK students were better performing the task than LPK students (Canham & Hegarty, 2010; Hegarty, Canham, & Fabrikant, 2010). Similar performance and processing differences were found in several other studies using eye tracking technology to analyze the processing behavior of participants with different knowledge levels while they solved various cognitive tasks (e.g., Amadieu, Van Gog, Paas, Tricot, & Mariné, 2009; Gegenfurtner, Lehtinen, & Säljö, 2011; Holmqvist, Andrà et al., 2011; Kaakinen, Hyönä, & Keenan, 2003; Van Gog, Paas, & Van Merriënboer, 2005).

Thus, similar processing differences between HPK and LPK students will also be assumed for the present cognitive task of solving MC items; that is, because domain knowledge in the form of schemas should optimize information processing. Therefore, HPK students are expected to focus more on information that is most relevant for correctly solving the task than LPK students. In the case of correctly solving MC items, this means that even though all answer options should be processed to some extent, the focus of attention should be more on the correct rather than on the incorrect answer options with increasing knowledge levels because, referring to gaze bias findings, paying more attention to an option is associated with a higher likelihood of choosing the option (e.g., Glaholt et al., 2009). Therefore, HPK students should pay more attention to correct answer options than LPK students. Moreover, given that HPK students should have automatized schemas for the target domain, they should be able to routinely process much of the information, which at least partly replaces the conscious cognitive processing of this information. As a result, they should be able to solve MC items quicker and with less effort than LPK students.

Existing eye tracking studies in the area of MC testing have already investigated the influence of different knowledge levels on MC item-solving, although with a more exploratory focus. A study from Tai, Loehr and Brigham (2006) investigated six students who solved MC items in three science domains for which they did or did not have high domain knowledge. A qualitative analysis revealed differences in students' scanpaths between their high and low prior knowledge domains. Moreover, studies by Andrà et al. (2009) and Holmqvist, Andrà et al. (2011) revealed that HPK students tended to show a more 'focused behavior' than LPK students when solving MC items in the domain of mathematics. Most of the time, especially HPK students compared smaller sets of answer options or the item stem. In contrast, LPK students more often demonstrated 'overview behavior', characterized by comparing larger sets of information units at a time. These first studies in the MC context primarily used the spatial information of gaze data to investigate differences in the location and distribution of attention related to prior domain knowledge. Furthermore, two recent studies took the temporal characteristics of gaze into account (Tang & Pienta, 2012; Tsai, Hou, Lai, Liu & Yang, 2012). They showed that students spent more time on options they chose than on options they rejected, which we would interpret as a first hint at a gaze bias effect when solving MC items. Furthermore, students successful in the MC task focused more on task-relevant information than students who failed to solve the task correctly (Tsai et al., 2012). Additionally, a study from Tang & Pienta (2012) showed that unsuccessful students inspected and revisited the item stem more frequently during the solution process than successful students.

Taken together, studies reported this far have found differences in MC item processing that are related to students' level of domain knowledge. However, besides having a strong exploratory character, most of these studies did not go into detailed quantitative analyses of the eye movement data. This type of analysis will be reported in the present paper. Moreover, in reported studies on knowledge-related differences in MC item processing, the MC test included visualizations such as complex graphics, mathematical formulas or diagrams, constituting an essential component in the solution process (cf., Saß, Wittwer, Senkbeil, & Köller, 2012). Hence, including such visualizations may lead to considerable differences in the processing of MC material in general. Additionally, items required higher order thinking with a focus on problem solving. As MC items that assess *knowledge* are often presented in an entirely *text-based* form in educational assessment, in the present research, students' processing of this specific type of MC items will be analyzed.

Apart from variations in the item format, to the best of our knowledge, the present study is the first to explicitly investigate *decision making* and, thus, the potential occurrence of the gaze bias effect in the applied and highly knowledge-related context of MC testing.

2 Research Questions

The main goal of the present research was to gain fine-grained insights into how students solve MC items as a function of their domain knowledge level and subjective preferences for answer options. By integrating insights from research on decision making (cf., gaze bias; Shimojo et al., 2003) and knowledge-related differences in cognitive processing (e.g., Sweller et al., 1998; Canham & Hegarty, 2010), we derived three hypotheses that will be presented in the following.

(1) Knowledge identification hypothesis. Because MC items are constructed to measure domain knowledge on an outcome level, we expect students with high domain knowledge (high prior knowledge; HPK) to achieve higher MC test scores (more correct solutions) while being more certain when responding (fewer preferred options) compared to students with low domain knowledge (low prior knowledge; LPK). Based on the assumption that HPK students, in contrast to LPK students, possess automatized schemas about the target domain (cf., Sweller et al., 1998), we expect HPK students to complete the test faster than LPK students (shorter response times). Furthermore, we expect HPK students to optimize information processing, which means that they should focus more on information that is most relevant for correctly solving the task. Therefore, HPK students should have higher relative fixation times on correct answer options than LPK students. Assuming that especially HPK students possess partial knowledge even when failing to choose the correct answer, we expect them to exceed LPK students' percentage of fixation time on correct answer options, also in cases of incorrect choices.

(2) Gaze bias hypothesis. As MC tests can be considered applied decision situations, relying on gaze bias findings (e.g., Glaholt & Reingold, 2011), we expect students to fixate longer on answer options with higher ratings of *subjective* preference. More specifically, we assume that total fixation times increase in a monotonic manner along distinct categories of increasing preference (i.e., non-attractive, attractive and chosen options). In the decision making literature, it is widely accepted that the gaze bias effect is strongly related to preference decisions (e.g., Glaholt & Reingold, 2011; Shimojo et al., 2003), though not exclusively (cf., Schotter et al., 2010). With this in mind, we assume that *subjective* preference is the driving force of the gaze bias effect, even in MC testing. Even though domain knowledge most probably influences the correctness of the options that students with different knowledge levels prefer, we expect the gaze bias towards preferred and chosen options to occur similarly for both HPK and LPK students. This is because the resulting subjective preference is decisive here, and not the source of the preference (e.g., domain knowledge).

Even though this assumption might seem to contradict the preceding hypothesis at first sight, they are actually perfectly compatible. Whereas the *knowledge identification* *hypothesis* assumes an attention bias towards correct answer options - especially for HPK students, the *gaze bias hypothesis* assumes a gaze bias towards subjectively preferred (and chosen) answer options for both HPK and LPK students. Thus, the focus of analysis on either *correct answer options* or on *subjective preferred options* determines the expected relation between fixation time and students' prior knowledge levels.

(3) Gaze bias consolidation hypothesis. Building on gaze likelihood analyses (cf., Shimojo et al., 2003), empirical studies showed that the gaze bias effect occurs mostly in the final part of the decision process (Glaholt & Reingold, 2009a, 2009b, 2011; Schotter et al., 2010; Shimojo et al., 2003). Along with these findings we expect the gaze bias to occur especially in the final phase of the decision process when solving MC items; that is, right before students' choice announcement. In particular, we expect an increase in attention towards the *chosen* options, where on the other hand we expect a decrease in attention towards *non-chosen* options in the final stage of the solution process. This pattern is expected for both HPK and LPK students.

Finally, we conducted an exploratory analysis of students' attention distribution across the whole decision making process in MC assessment, taking students' subjective preferences for answer options into account. In particular, we analyzed whether HPK and LPK students differed in their overall time-course of decision making when solving MC items.

3 Method

3.1 Participants and Design

Participants were 26 students from the University of Kiel in Germany. All students were native German speakers with normal or corrected to normal vision. Two distinct groups of participants were recruited due to their respective fields of study. The 'HPK group' comprised only master students in psychology (n = 14; 57 % female; M_{age} = 24.4 years, SD_{age} = 1.91) since the MC test used in the present study assessed domain knowledge in neurology. The 'LPK group' comprised students in economics or law (n = 12; 66 % female; M_{age} = 23.4 years, SD_{age} = 2.19). Originally, 30 students participated in the study but data from four students (14 %; 1 HPK and 3 LPK students) had to be excluded from the analysis due to poor eye tracking data quality or data loss. Hence, the total sample consisted of N = 26.

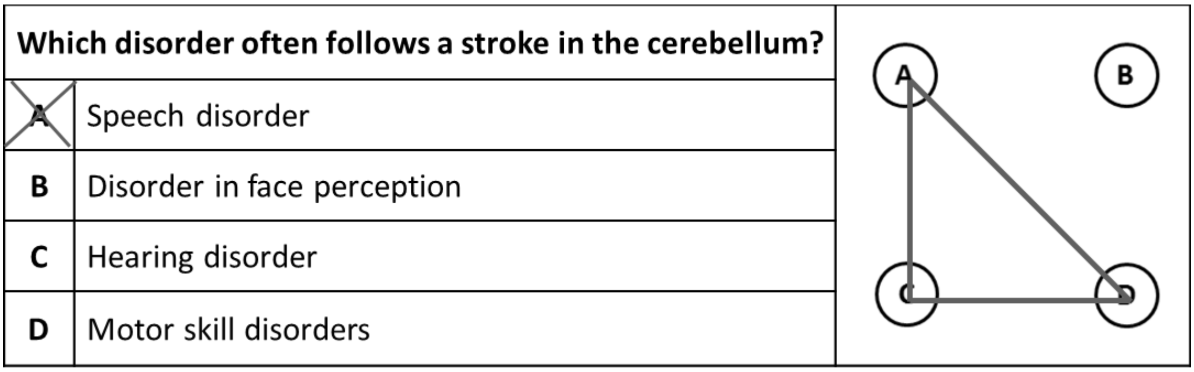
3.2 Materials and Measures

Multiple-choice knowledge test. The MC test we applied in the present study (Thoma, Dalehefte & Köller, 2014) consisted of 21 items in the domain of biological psychology regarding the topic of the brain. Every item comprised a short item stem formulated as a question and four short answer alternatives that were displayed below the item stem. For reasons of comparison, all item stems and answer alternatives were not longer than one sentence. All items had the same MC format, meaning that in each of the items only one out of four answer alternatives was correct (the remaining three alternatives were distractors).

Item development was oriented along psychology study regulations concerning contents in the subjects of biological psychology and neurosciences while we also adhered to MC item writing guidelines (e.g., Haladyna et al., 2002; Haladyna, 2004). In addition, six independent experts in the field of neurosciences, biology and neurology reviewed the test and verified item contents. Referring to the cognitive taxonomy of Bloom, Engelhart, Furst, Hill and Krathwohl (1956), our items pertain to the domain of *knowledge* as students mostly had to recall information without performing higher cognitive transformations or problem solving. It was not possible to solve items by logical thinking or general knowledge alone; hence, a certain expertise in the neuroscience domain was needed to achieve high test scores. Students were given one point for each correct answer to a test item (maximum test score of 21 points). To evaluate psychometrical test characteristics, the MC items were pretested in a paper-pencil version. Participants of the pretest were N = 377 students with different fields of study, and thus with high or low prior knowledge in the domain of neuroscience. As a high prior knowledge subsample group ($n = 149$), medical and psychology students participated in the test calibration study. The pretest sample was comparable to participants that took part in the present study. Item difficulty ranged between .17 and .86, while most items had a moderate difficulty ($M = .55$), allowing for good differentiation between students. Item discrimination values ranged from .30 to .75 ($M = .53$). Internal consistency as measured by the Kuder-Richardson Formula-20 was satisfying with $\alpha = .88$. Test validity was supported by the test scores in the students' subsamples. Medical and psychology students had significantly higher test scores than students in other fields of study ($M_{HPK} = 15.7$; $SD_{HPK} = 4.9$; $M_{LPK} = 8.4$; $SD_{LPK} = 3.0$; $t(219.8) = 16.2$, $p < .001$, d = 1.85).

Verbal stimulus characteristics. In the present MC test, all answer options were constructed to be highly comparable in their verbal characteristics. Accordingly, aggregated across items correct answer options and distractors were highly similar both concerning their mean number of words ($M_{Correct}$ = 2.48, $SD_{Correct}$ = 1.57; $M_{Distractors}$ = 2.24, $SD_{Distractors}$ = 1.57), concerning their mean number of characters ($M_{Correct}$ = 17.14, $SD_{Correct}$ = 12.76; $M_{Distractors}$ = 15.97, $SD_{Distractors}$ = 12.10), and concerning their familiarity ($M_{Correct}$ = 14.57, $SD_{Correct}$ = 3.26; $M_{Distractors}$ = 14.80, $SD_{Distractors}$ = 3.39), which was operationalized via the frequency of the words in the German language[1].

Preference rating system. To evaluate students' preferences for answer options in the MC test, we constructed a paper-pencil answer system (MC-ASYS) allowing for a preference differentiation on three levels. In that system every item is presented once again after the test with an additional field next to the item (on the right side) in which the answer options are symbolized by a circled letter and arranged in an equidistant square (see Figure 1).



Which disorder often follows a stroke in the cerebellum?	
~~A~~	Speech disorder
B	Disorder in face perception
C	Hearing disorder
D	Motor skill disorders

Figure 1. Structure of the paper-pencil preference system (MC-ASYS), illustrated by an example item (translated from German) with a hypothetic rating: The student *chose* answer option A in the computer MC test (= chosen option), but also preferred options C and D (= attractive options) while option B could be excluded with high confidence (= non-attractive option).

---

[1] Word frequency parameters were taken from the homepage 'http://wortschatz.uni-leipzig.de/' [*last access*: 12.12.2013], which provides information about the frequency class of every German word in reference to the most frequent word 'der' (translated as 'the'). A frequency class of 'x' for a certain word means that 'der' is two, raised to the exponent x (2^x) times more common than the searched word. As words in the test were extremely heterogeneous (either very seldom like 'hippocampus' or very common like 'and'), we skipped all common words with a frequency class less than or equal to 10 and calculated means only for more uncommon words to prevent a bias in the calculation of means. We chose this procedure, as the uncommon words probably have the most important impact on fixation duration parameters.

Students were instructed to cross out the answer option they actually chose in the previous MC test (in order to control for their memory accuracy, which was very high (*M* = 95 % correct; *SD* = 0.05). Furthermore, in the right field they were asked to connect all answer options that they could not exclude as incorrect. These options are interpreted as preferred. Based on that rating, a discrimination between options that are preferred and chosen (3 points; i.e., *chosen*), preferred but not chosen (2 points; i.e., *attractive*) and not preferred at all (1 point; i.e., *non-attractive*) becomes possible. The MC-ASYS was applied to all 21 items. In total, students could mark all 84 answer options of the test (maximum uncertainty) indicating that they could not exclude any option, while at least 21 options had to be chosen (this would reflect maximum certainty).

Control variables. We controlled for the cognitive and spatial abilities of HPK and LPK students as well as for their test-taking motivation, since these factors might influence MC test processing aside from prior knowledge in the test domain. To assess general cognitive abilities, we administered the *N2* subtest of the KFT 4-12+R (Heller & Perleth, 2000). To measure spatial abilities, the *Paper Folding (N3)* subtest of the KFT 4-12+R was administered (Heller & Perleth, 2000). Students' current motivation to engage in the test was assessed via a 4-point Likert scale comprising six items (e.g., ”I will do my very best in this test.”).

3.3 Apparatus

Test items were presented on a 22-inch screen with a 1680x1050 pixel resolution, using the software Experiment Center 3.1 from SensoMotoric Instruments (SMI; Teltow, Germany). Every item was presented on a single page on the screen. While working on the test participants were sitting in front of the screen at a distance of approximately 70 centimeters. This resulted in an approximate font size of 0.7 degrees visual angle and a vertical stimulus size of about 16-19 degrees per item (top first line to bottom last line).

Participants' eye movements were recorded using a video-based eye tracking system (SMI iView X™ RED; 120Hz sampling rate) and the corresponding SMI Software iView X™. The system was calibrated using an animated 8-point calibration image and subsequent validation. Calibration accuracy was below 0.65 degrees of visual angle for both x and y coordinates for all participants (range: 0.14 to 0.64; M_x = 0.43, SD_x = 0.13; M_y = 0.35, SD_y = 0.11).

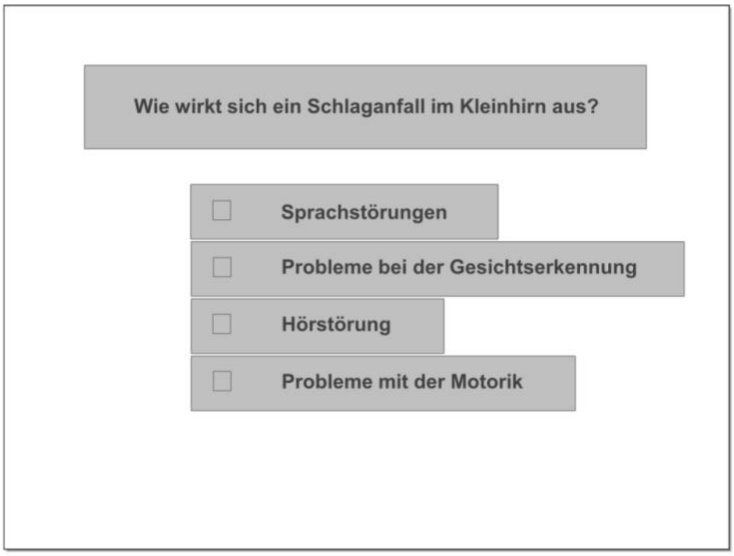
3.4 Procedure

Students were tested in single sessions. In the beginning, students answered paper-pencil questionnaires about demographics and their current test-taking motivation. Afterwards, they were familiarized with the procedure and the eye tracking system. Students completed a short example MC test with four easy-to-solve items to familiarize them with how to provide answers to the items by using the mouse to click on a field next to an answer option, and how to go on to the next question by pressing an arrow key. Students were informed in a short standardized text that they would not be able to return to an earlier question and had no time constraints for the task. Furthermore, they were instructed to choose exactly *one* answer to each item and to guess in case of doubt. The eye tracking system was calibrated before the MC test started. After finishing the test, students completed the spatial and cognitive ability tests and filled out the preference rating system (MC-ASYS; see Figure 1). It is noteworthy that additional data, namely cued retrospective think aloud protocols, questionnaires and another test regarding MC test-wiseness were collected after the MC test. We will not report these data here as they are not central with respect to answering the present research questions.

3.5 Eye Movement Data Pre-Processing

Eye movement recordings were analyzed using a dispersion-based algorithm implemented in the BeGaze™ software, version 3.0 from SMI. A fixation was detected when eye movements lasted for at least 80 milliseconds at a position with a maximum dispersion of 100 pixels. For each item, five separate rectangular 'areas of interest' (AoIs) were placed encompassing the question and each of the four answer options (see also Figure 2). A margin was added to the letters in the AoIs to account for data inaccuracy and failings of the participants to directly look at the text.

Mean AoIs sizes were comparable between correct answer options ($M_{Correct}$ = 66124 pixels; $M_{Correct}$ = 3.74 % stimulus coverage) and incorrect answer options ($M_{Distractors}$ = 63164 pixels; $M_{Distractors}$ = 3.58 % stimulus coverage). The question had a mean AoI size of $M_{Question}$ = 180276 pixels and covered about $M_{Question}$ = 10.2 % of the total stimulus. All reported fixation times refer to these five stimulus areas (per item); fixations on regions without visual information (white space) were excluded from analysis. Furthermore, eye movements that occurred after choosing an answer option (indicating that the decision process is terminated) were removed from the analysis using a self-programmed algorithm.


Wie wirkt sich ein Schlaganfall im Kleinhirn aus?

☐ Sprachstörungen
☐ Probleme bei der Gesichtserkennung
☐ Hörstörung
☐ Probleme mit der Motorik

Figure 2. Example of one original MC test item (black writing on white screen) in German language (translation can be taken from Figure 1 as this example shows the identical item) with overlying AoI drawings (five grey rectangles).

We used total fixation time as gaze data. Fixation time is defined as the sum of all consecutive fixations on an AoI, indicating the time of attending on this area, while total fixation time cumulates the duration of all AoI fixations from task onset until task end (cf., Holmqvist, Nyström et al, 2011). To test our hypotheses, we first analyzed fixation times for every item and every person, computing means on a within-person level. For further analyses on a group level, these intra-individual means are used as data.

To test the temporal gaze bias consolidation hypothesis, we programmed a Matlab® algorithm to divide the time course of the decision process into ten equal time intervals for every item and every person according to the five AoIs (item stem, 4 answer options). Therefore, we were able to conduct an adapted gaze likelihood analysis (e.g., Shimojo et al., 2003) with which we could compare the time-course of item processing for all persons and items. For every time interval, total fixation time was determined for all five AoIs, while answer options were subsequently classified with respect to ratings of preference. Given that the frequencies with which the preference categories *(non-attractive, attractive,* and *chosen)* were chosen differed in the MC-ASYS rating, we divided the fixation time data for each preference category (e.g., attractive options) by the frequency with which this category was chosen. This weighting procedure was necessary to allow for an unbiased comparison of fixation time data as a function of the different preference ratings in each time interval. This procedure was first carried out on a person level, calculating within-means for all items, and then these means were further analyzed on a group level for HPK and LPK students. All statistical analyses were conducted with IBM® SPSS® statistics (Version 19) software.

4 Results

The analysis is structured according to the hypotheses. Prior to testing the hypotheses, control variables were analyzed to determine whether HPK and LPK students were comparable with regard to their prior abilities and motivation for the test. Descriptive values are shown in Table 2.1. Three *t*-tests revealed no significant (two-tailed) group differences between HPK and LPK students regarding their *general cognitive abilities*, *t*(24) = 0.91, *p* = .37, *spatial abilities*, *t*(24) = 0.85, *p* = .40, and *test taking motivation*, *t*(24) = 1.53, *p* = .14.

Table 1. Means and standard deviations for prior abilities, test taking motivation, preference ratings and MC test outcome as a function of students' level of domain knowledge.

	HPK Students *N* = 14; *M* (*SD*)	**LPK Students** *N* = 12; *M* (*SD*)	***p* value, two-tailed** (*t-value*$_{df}$)
Spatial abilities (min. = 0, max.= 10)	6.57 (*2.74*)	5.75 (*2.05*)	.40 (*0.85*$_{24}$)
Cognitive abilities (min. = 0, max.= 25)	18.29 (*3.87*)	17.23 (*3.19*)	.37 (*0.91*$_{24}$)
Test-taking motivation (min. = 0, max.= 24)	19.07 (*3.07*)	20.83 (*2.75*)	.14 (*1.53*$_{24}$)
Preference rating score[1] (min. = 0, max. = 63)	19.00 (5.78)	28.25 (9.96)	.007 (*2.94*$_{24}$)
MC test score (min. = 0, max.= 21)	15.21 (*2.83*)	9.50 (*2.32*)	<.001 (*5.57*$_{24}$)

Note. [1] Reported preference scores refer to options that students could not exclude in addition to the one option they had to select in every item. Therefore, one point was subtracted for each item, resulting in a maximum preference score of 63 (originally = 21 items multiplied by 4 options = 84 points as total score). High values on this scale indicate substantial uncertainty in responding.

4.1 Knowledge Identification Hypothesis

Due to the procedure of recruiting participants according to their respective fields of study for the HPK and the LPK group, differences in the test scores were used to validate the domain knowledge status of the groups.[2] A t-test revealed that HPK students answered significantly more questions correctly than LPK students ($M_{HPK} = 15.21$, $SD_{HPK} = 2.83$; $M_{LPK} = 9.5$, $SD_{LPK} = 2.32$; $t(24) = 5.57$, $p < .001$, $d = 3.00$), accounting for the expected difference in domain knowledge levels.

The total number of answer options that could not be excluded as incorrect (for the entire test) in the MC-ASYS rating were significantly higher for LPK students ($M_{HPK} = 19.00$, $SD_{HPK} = 5.78$; $M_{LPK} = 28.25$, $SD_{LPK} = 9.96$; $t(24) = 2.94$, $p = .007$, $d = 1.23$). Thus, as expected LPK students considered more distractors as being potential correct solutions to the question than HPK students did, accounting for HPK student's higher certainty in responding. Nevertheless, even though LPK students excluded the correct answer option significantly more often than HPK students ($M_{HPK} = 4.42$, $SD_{HPK} = 1.8$; $M_{LPK} = 2.36$, $SD_{LPK} = 2.6$; $t(24) = 2.42$, $p = .023$, $d = 0.94$), both groups revealed that they possess a certain amount of partial knowledge, as both student groups seldom classified the correct answer as being non-attractive and therefore seldom excluded it as the potential correct answer option.

Furthermore, HPK and LPK students significantly differed in the mean duration (in milliseconds) they needed to complete an MC item ($M_{HPK} = 9584$, $SD_{HPK} = 1238.3$; $M_{LPK} = 13886$, $SD_{LPK} = 4171.3$; $t(12.7) = 3.45$, $p = .001$, $d = 1.40$)[3].

According to the second part of the hypothesis, namely that HPK students spend more time on correct answers than LPK students, we calculated the fixation time students spent on the correct option relative to the time they spent on all answer options (per item). These relative fixation times were averaged on a person level. Results indicate that HPK students spent significantly more time on correct answer options compared to LPK students overall ($M_{HPK} = 0.36$, $SD_{HPK} = 0.04$; $M_{LPK} = 0.29$, $SD_{LPK} = 0.02$; $t(24) = 5.70$, $p < .001$,

---

[2] To adjust α-levels for multiple comparisons in the following five t-Tests, we applied the 'Holm-Bonferroni Procedure' (Holm, 1979) to prevent α-inflation. With a global α-level of $\alpha_g = .05$, tests were significant as reported in the text with the following calculated local α-levels: $\alpha_1 < .008$; $\alpha_2 < .01$; $\alpha_3 = .0125$; $\alpha_4 = .016$; $\alpha_5 < .025$; $\alpha_6 < .05$).

[3] T-test results are reported with corrected degrees of freedom if the assumption of homogeneity of variance was violated.

d = 2.42), and even when answering incorrectly (M_{HPK} = 0.23, SD_{HPK} = 0.06; M_{LPK} = 0.19, SD_{LPK} = 0.03; *t*(24) = 2.40, *p* = .024, *d* = 0.95). We furthermore correlated the mean of relative fixation time on correct answer options for the whole MC test with students' final test score, which resulted in a high positive correlation for HPK students (*r* = .848, *p* < .001), and a moderate but non-significant correlation for LPK students (*r* = .383, *p* = .22).

4.2 Gaze Bias Hypothesis

To test the hypothesis that HPK and LPK students show a comparable gaze bias effect, reflected by a monotonic increase of total fixation time on answer options with higher rated preference (see Figure 3), we conducted a 2x3 mixed analysis of variance (ANOVA[4]) with *domain knowledge* as between-subjects factor, *preference* as within-subject factor and *total fixation time* as dependent measure.

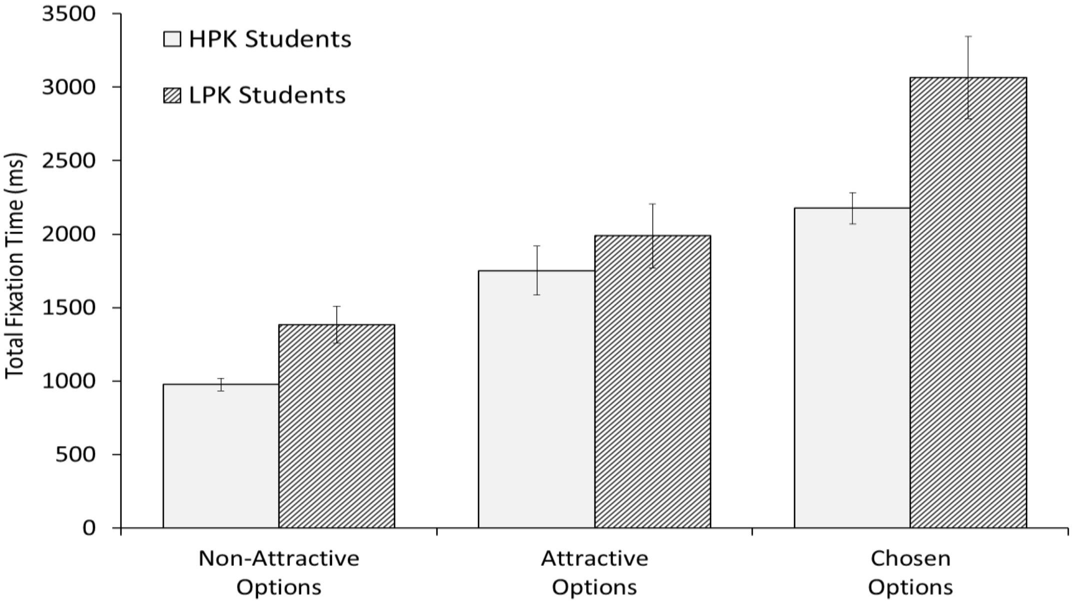


Figure 3. Averaged total fixation times and standard errors (in milliseconds) on MC answer options (AoIs) according to students' subjective preference ratings during the entire item solution process, displayed separately for groups of high prior knowledge (HPK) and low prior knowledge (LPK) students.

---

[4] ANOVA results are reported with Greenhouse-Geisser corrected degrees of freedom if the assumption of sphericity was violated.

The ANOVA revealed significant main effects of *domain knowledge*, $F(1, 24) = 6.56$, $p = .017$, $\eta^2_p = .215$, and *preference*, $F(2, 48) = 86.98$, $p < .001$, $\eta^2_p = .784$. Furthermore, a significant interaction between both factors, $F(2, 48) = 4.80$, $p = .013$, $\eta^2_p = .167$, was detected. Descriptive values are shown in Table 2.2.

Table 2. Averaged total fixation times on MC answer options (AoIs) to correspondent participant-rated levels of subjective preference as a function of students' level of domain knowledge (high prior knowledge (HPK) vs. low prior knowledge (LPK) students).

Total Fixation Time	Non-Attractive Options *M (SD)*	Attractive Options *M (SD)*	Chosen Options *M (SD)*	All Options *M (SD)*
HPK Students (*N* = 14)	976.7 (156.6)	1752.1 (612.9)	2176.4 (395.3)	1635.1 (655.1)
LPK Students (*N* = 12)	1382.8 (432.7)	1989.4 (754.8)	3066.6 (970.8)	2146.3 (1016.3)

Note. Total fixation times are reported in milliseconds.

As the main effect for *domain knowledge* was significant, we conducted Bonferroni adjusted post hoc comparisons between HPK and LPK students at each *preference-level* to further explore the interaction.

Results showed that LPK students attended longer than HPK students on non-attractive options, $F(1, 24) = 10.7$, $p = .003$, $\eta^2_p = .309$, and on chosen options, $F(1, 24) = 9.9$, $p = .004$, $\eta^2_p = .292$, while no difference was found for attractive options, $F(1, 24) = 0.8$, $p = .385$, $\eta^2_p = .032$, accounting for the significant interaction of *domain knowledge* and *preference*. As the main effect of *preference* was also significant, we conducted Bonferroni adjusted post hoc comparisons to further explore the gaze bias effect (more attention on more preferred answer options) for HPK and LPK students separately. These analyses showed significant differences between all three levels of preference ($p < .05$) for both of the groups, while means indicated a monotonic increase in fixation times along increasing option preference.

4.3 Gaze Bias Consolidation Hypothesis

We tested the hypothesis that the gaze bias effect occurs especially in the final stage of the decision process. Specifically, we expected an increase in attention towards the chosen options and a decrease in attention towards the non-chosen options right before choice announcement for both HPK and LPK students.

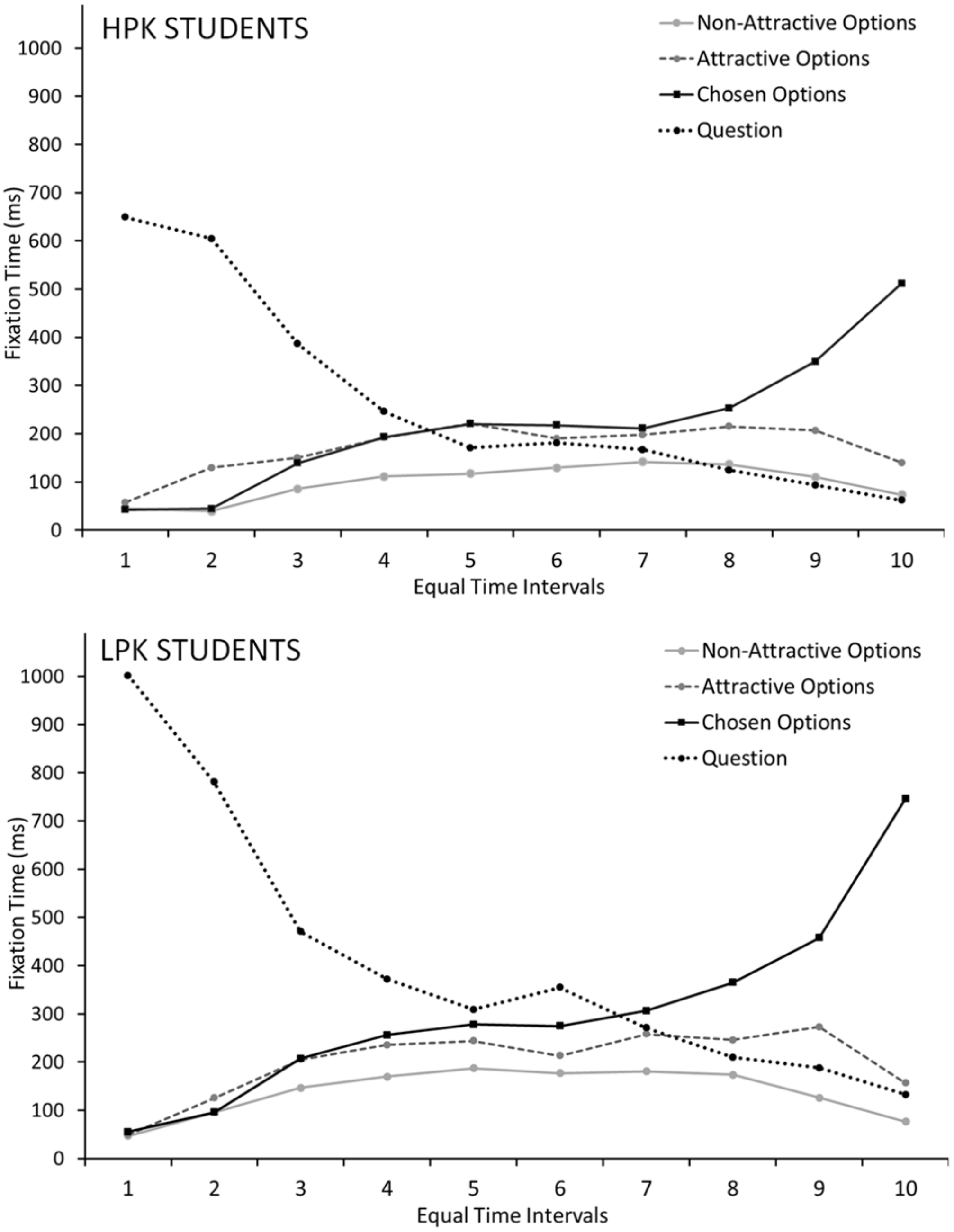


Figure 4. High prior knowledge (HPK) and low prior knowledge (LPK) students' total fixation times across ten equal time intervals for the question, the chosen options, the attractive options, and the non-attractive options (averaged across items for the separate groups).

Prior to testing this hypothesis we analyzed how students generally distributed their attention across the whole item-solving process, and whether the patterns of attention allocation differed between HPK and LPK students. Therefore we conducted a mixed 2x4x10 ANOVA with *domain knowledge* as between-subjects factor and *preference* and *time interval* as within-subject factors. This analysis revealed significant main effects for the factors *domain knowledge*, $F(1, 24) = 11.01$, $p = .003$, $\eta^2_p = .315$, *preference*, $F(2, 53) = 7.83$, $p < .001$, $\eta^2_p = .822$ and *time interval*, $F(5, 119) = 20.96$, $p < .001$, $\eta^2_p = .466$. Furthermore, interactions between *domain knowledge* and *preference*, $F(2,53) = 7.83$, $p = .001$, $\eta^2_p = .246$, *domain knowledge* and *time interval*, $F(5, 119) = 2.34$, $p = .046$, $\eta^2_p = .089$, *preference* and *time interval*, $F(9, 119) = 114.6$, $p < .001$, $\eta^2_p = .827$, as well as the three-way interaction between *domain knowledge*, *preference* and *time interval*, $F(9, 205) = 3.62$, $p < .001$, $\eta^2_p = .131$, were significant. To break down the three-way interaction, we analyzed interactions between *domain knowledge* and *time interval* separately for the question and the preference categories (i.e., chosen options, attractive options, non-attractive options) in four 2x10 ANOVAs. This was, aside from the intention to explain the significant three-way interaction, also an exploratory analysis with the goal to quantitatively describe the item-solving process for HPK and LPK students on a fine-grained level.

To avoid too many numbers in the text, we report the detailed ANOVA results in Table 2.3 and only selected *p*-values in the text. In Figure 4 and 5 students' total fixation times across ten equal time intervals are displayed for the question, the chosen options, the attractive options, and the non-attractive options (averaged across items for both groups).

The 2x10 ANOVAs revealed significant main effects of *domain knowledge* for the question, the chosen and the non-attractive options (all $p_s < .01$), but not for the attractive options ($p = .26$). All ANOVAs revealed main effects of *time interval* (all $p_s < .001$), while interactions between *domain knowledge* and *time interval* were significant for the question, the chosen and the non-attractive options (all $p_s < .01$), but not for the attractive options ($p = .56$). To further explore the significant interactions between *domain knowledge* and *time interval*, we conducted the respective repeated contrasts, of which we report only the *p*-values of the significant findings to avoid too many numbers in the text (see Table 2.4 for all contrasts).

Table 3. Parameters and results from four mixed ANOVAs (for the question, chosen options, attractive options and non-attractive options) with ‘domain knowledge’ as between-subjects factor and ‘time interval’ as within-subject factor.

	Domain Knowledge					Time Interval					Domain Knowledge *Time Interval				
	F	*df*	*MSE*	*p*	η^2_p	*F*	*df*	*MSE*	*p*	η^2_p	*F*	*df*	*MSE*	*p*	η^2_p
Question	14.9	1, 24	85780.1	**.001**	.383	186.2	4, 97	18091.1	**<.001**	.886	5.46	4, 97	18091.1	**<.001**	.185
Chosen Options	9.5	1, 24	50147.8	**.005**	.284	114.1	5, 120	11269,9	**<.001**	.826	3.69	5, 120	11269,9	**.004**	.133
Attractive Options	1.3	1, 24	45678.6	.260	.052	17.4	5, 126	9144.2	**<.001**	.420	0.80	5, 126	9144.2	.560	.032
Non-Attractive Options	9.4	1, 24	10478.5	**.005**	.282	33.3	9, 216	1361.9	**<.001**	.581	2.90	9, 216	1361.9	**.003**	.108

Note. Results are reported with Greenhouse-Geisser corrected degrees of freedom if the assumption of sphericity was violated.

Table 4. Parameters and results for repeated contrasts for the interaction of the factors ‘domain knowledge’ and ‘time interval’ (for the question, chosen options, attractive options and non-attractive options).

	Question			Chosen Options			Attractive Options			Non-Attractive Options		
	$F_{(1,24)}$	*p*	η^2_p	$F_{(1,24)}$	*p*	η^2_p	$F_{(1,24)}$	*p*	η^2_p	$F_{(1,24)}$	*p*	η^2_p
1 vs. 2	5.8	**.024**	.196	3.8	.061	.138	< 0.1	.926	.000	8.6	**.007**	.265
2 vs.3	6.7	**.016**	.218	0.3	.619	.010	3.6	.070	.131	< 0.1	.805	.003
3 vs. 4	0.9	.343	.037	< 0.1	.887	.001	0.1	.738	.005	< 0.1	.884	.001
4 vs. 5	0.2	.689	.006	< 0.1	.885	.001	0.3	.599	.012	0.5	.468	.022
5 vs. 6	1.0	.328	.040	< 0.1	.980	.000	< 0.1	.989	.000	1.8	.196	.069
6 vs. 7	2.3	.139	.089	0.9	.342	.038	0.9	.361	.035	0.2	.656	.008
7 vs. 8	0.5	.496	.019	0.2	.693	.007	0.7	.424	.027	< 0.1	.929	.000
8 vs. 9	0.1	.758	.004	< 0.1	.912	.001	0.9	.360	.035	2.0	.167	.078
9 vs.10	0.9	.357	.035	8.3	**.008**	.257	2.7	.114	.101	0.5	.448	.020

Note. Degrees of freedom are *F*(1, 24) for all reported repeated contrasts.

Table 5. Parameters and results for repeated contrasts for the factor 'time interval' (for the question, chosen options, attractive options and non-attractive options), conducted separately for high prior knowledge (HPK) and low prior knowledge (LPK) students.

Intervals HPK Students	Question			Chosen Options			Attractive Options			Non-Attractive Options		
	$F_{(1,13)}$	p	η^2_p	$F_{(1,13)}$	p	η^2_p	$F_{(1,13)}$	p	η^2_p	$F_{(1,13)}$	p	η^2_p
1 vs. 2	1.0	.328	.074	0.1	.819	.004	6.7	.022	.341	0.4	.529	.031
2 vs.3	107.4	.000	.892	32.9	<.001	.716	1.2	.294	.084	49.5	<.001	.729
3 vs. 4	20.2	.001	.609	7.5	.017	.366	3.6	.081	.216	13.4	.003	.506
4 vs. 5	12.5	.004	.491	1.4	.259	.097	1.4	.251	.099	0.6	.464	.042
5 vs. 6	0.4	.523	.032	<0.1	.910	.001	1.9	.189	.129	2.1	.175	.137
6 vs. 7	0.5	.512	.034	0.1	.811	.005	0.1	.767	.007	1.6	.234	.107
7 vs. 8	6.2	.027	.322	2.8	.120	.176	0.6	.448	.045	0.2	.654	.016
8 vs. 9	5.2	.039	.288	15.6	.002	.546	0.1	.716	.011	6.8	.022	.344
9 vs.10	3.9	.071	.229	26.5	<.001	.671	24.2	.000	.651	11.9	.004	.497

Intervals LPK Students	**Question**			**Chosen Options**			**Attractive Options**			**Non-Attractive Options**		
	$F_{(1,11)}$	p	η^2_p	$F_{(1,11)}$	p	η^2_p	$F_{(1,11)}$	p	η^2_p	$F_{(1,11)}$	p	η^2_p
1 vs. 2	13.4	.004	.549	4.9	.049	.307	6.2	.030	.362	8.1	.016	.425
2 vs.3	109.4	.000	.909	14.1	.003	.561	10.2	.008	.482	6.5	.027	.371
3 vs. 4	11.1	.007	.501	2.6	.133	.193	1.4	.257	.115	1.3	.274	.107
4 vs. 5	7.3	.021	.389	0.7	.416	.061	<0.1	.803	.006	1.7	.222	.132
5 vs. 6	1.9	.199	.145	<0.1	.915	.001	0.9	.354	.078	0.5	.515	.040
6 vs. 7	3.8	.077	.257	1.2	.301	.097	2.0	.187	.153	0.1	.823	.005
7 vs. 8	9.0	.012	.450	3.2	.103	.223	0.2	.682	.016	0.2	.678	.016
8 vs. 9	0.7	.422	.060	14.0	.003	.560	0.7	.411	.062	16.4	.002	.589
9 vs.10	7.6	.019	.409	94.9	.000	.896	16.8	.002	.605	10.9	.007	.489

Note. Degrees of freedom are $F(1, 24)$ for all reported repeated contrasts.

Repeated contrasts revealed that fixation times on the *question* developed differently for HPK and LPK students between the first and the second time interval (*p* = .02) as well as between the second and third time interval (*p* = .02). Fixation times developed differently for HPK and LPK students only between the ninth and the tenth time interval for *chosen* options (*p* = .01), and only between the first and the second time interval for *non-attractive* options (*p* = .01). Thus, there were not many differences between the slopes of the fixation time curves across the whole item-solving process between HPK and LPK students (in 4 out of 36 time intervals); differences were only found in the beginning or at the end of the time course. This pattern of results is corroborated by the similar time courses of the fixation time curves between HPK and LPK students as displayed in Figure 4 and 5.

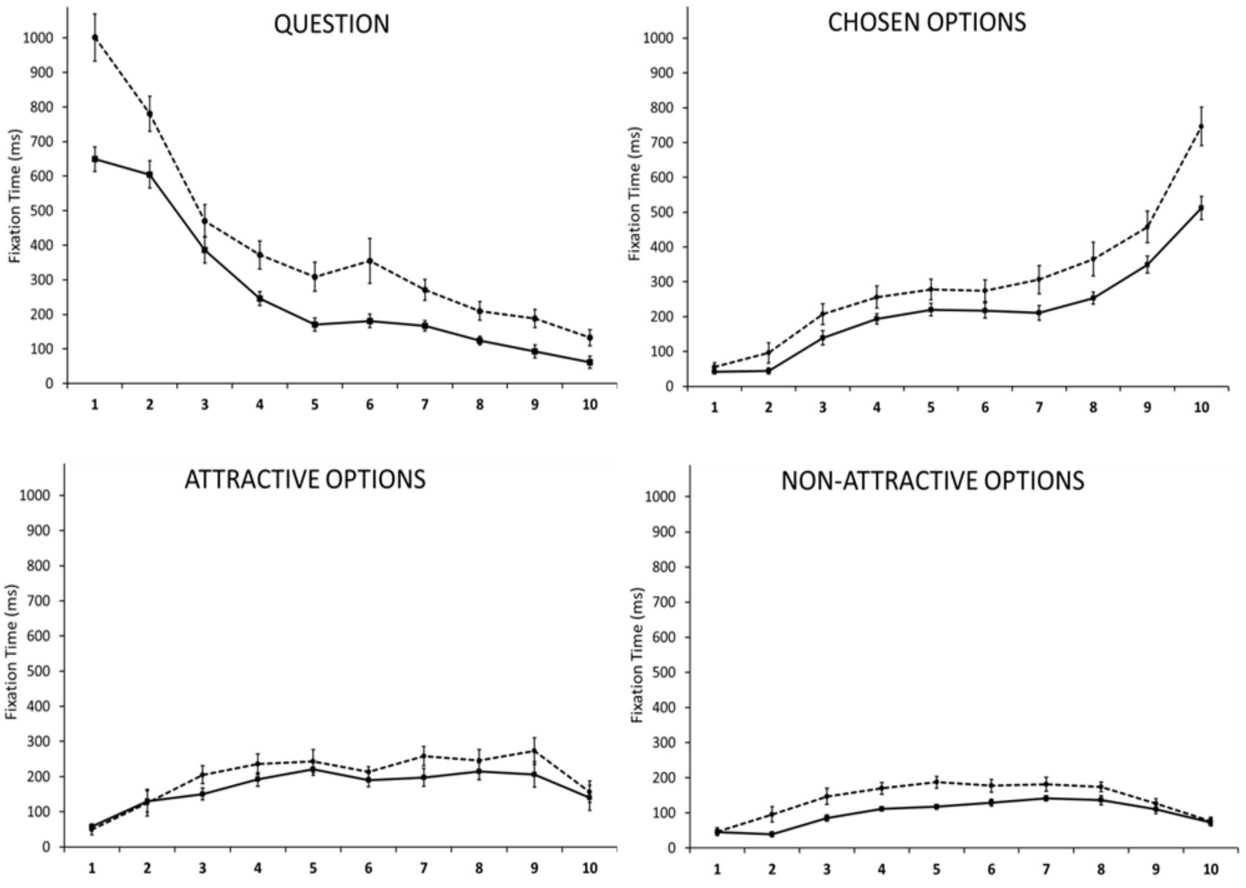


Figure 5. Time course of high prior knowledge (HPK; solid line) and low prior knowledge (LPK; dashed line) students' total fixation times across ten equal time intervals displayed separately for the question, the chosen options, the attractive options, and the non-attractive options (with standard error bars).

To test our hypothesis of a relative increase in attention towards the chosen options especially in the final stages of the decision process for both HPK and LPK students (i.e., temporal gaze bias consolidation hypothesis), we further conducted repeated contrasts for the factor *time interval* for HPK and LPK students separately (see Table 2.5 for all contrasts).

HPK and LPK students' fixation times on *chosen* options increased monotonically across the whole solution process. Moreover, for both HPK and LPK students fixation times increased from the eighth to the tenth time interval (both $p_s < .01$), and thus as expected between the last three time intervals of the decision process. Fixation times for *attractive* options did not differ between the eighth and the ninth time interval (both $p_s > .40$), and decreased between the ninth and the tenth time interval for both HPK and LPK students (both $p_s < .01$). Fixation times for *non-attractive* options decreased from the eighth to the tenth time interval for both HPK and LPK students (all $p_s < .05$). These data support our hypothesis that the gaze bias effect occurs mostly in the final stage of the decision process, and in a comparable manner for both prior knowledge groups.

5 Discussion

In this eye tracking study we investigated HPK and LPK students' MC item-solving processes on a fine-grained level. By taking theoretical assumptions and empirical findings from the research areas of knowledge-related differences in cognitive processing and decision making into account, we derived hypotheses about students' choice behavior in the educational context of MC assessment.

According to the *knowledge identification hypothesis,* we first expected HPK students to have higher test scores while being more certain and faster in responding to the test items than LPK students. Supporting this hypothesis, HPK students solved more items correctly, considered fewer answer options as potentially being correct (reflecting higher certainty), and were faster in completing the test than LPK students. These results can be interpreted as confirming the assumption that persons with high domain knowledge possess well-organized and automatized schemas in long-term memory, allowing them to efficiently process information in their knowledge domain (cf., Sweller et al., 1998). Moreover, domain knowledge in the form of schemas was expected to enable HPK students to optimize their processing of MC items by focusing more on information that was most relevant with regard to solving the items correctly; that is, by paying closer attention to the correct answer. Results revealed that the percentage of time spent fixating on correct answer options was indeed considerably higher for HPK than for LPK students. Consequently, HPK students fixated much less on incorrect distractors than LPK students. This was even true when HPK students failed to choose the correct answer to a test item, suggesting that, as expected, HPK students possessed more (partial) knowledge than LPK students, even for those test items that they solved incorrectly.

In line with other gaze bias findings (e.g., Glaholt & Reingold, 2011; Shimojo et al., 2003), in the present study, increased attention to correct answer options was also associated with choosing them more frequently. This is reflected in the correlation between the mean time spent on correct answer options and the over-all test score, which was high and significant for HPK students. Thus, the amount of attention paid to correct answer options (i.e., fixation time) was a reliable indicator for the MC test performance of HPK students and enabled a prediction of their test scores. Moreover, since HPK students have been found to pay more attention to correct answer options than LPK students, the amount of attention spent on correct answer options in the present MC test was indicative of the students' level of domain knowledge.

For LPK students, the correlation between the mean time spent on correct answer options and the overall test score was moderate but not significant. This might be due to their insufficient domain knowledge, possibly leading to a more random attention distribution on answer options as well as to guessing behavior, which in turn produces unsystematic variance and low reliability values. This may explain why there was a reliable connection of eye movements and MC test results only for HPK students, for whom the test was constructed in the first place. To determine whether HPK students' gaze and test scores are tightly coupled across different test domains and test formats, further studies are needed to provide a larger empirical basis.

Following the *gaze bias hypothesis*, we expected both HPK and LPK students to show a gaze bias effect, and thus, to fixate longer on answer options with higher ratings of subjective preference (regardless of their objective correctness). In agreement with this hypothesis, eye movement data revealed a monotonic increase in fixation times as a function of higher preference ratings for answer options on a three-categorical rating scale (*non-attractive options* vs. *attractive options* vs. *chosen options*). Thus, the present findings replicate well established findings from basic research on decision making in the context of MC items, accounting for the high stability of the gaze bias effect across various decision paradigms (e.g., Bird et al., 2012; Flowe, 2011; Flowe & Cottrell, 2011; Glaholt & Reingold, 2011; Pieters & Warlop, 1999; Shimojo et al., 2003).

Moreover, the pattern of the gaze bias effect was comparable for both HPK and LPK students, supporting the assumption that the gaze bias not only occurs for students with a certain level of domain knowledge in MC item-solving, but rather for students with various levels of domain knowledge. However, HPK students paid almost as much attention to attractive options as to chosen options, whereas LPK students paid much more attention to chosen than to attractive options (see Figure 3). This is reflected in the significant interaction between the factors *domain knowledge* and *preference*.

This interaction might go back to the fact that different sources for subjective preferences in MC item-solving depend on the level of prior domain knowledge. Whereas subjective preference (or perceived attractiveness) is probably based on existing domain knowledge for HPK students, it might be based on informed guessing for LPK students. Therefore, HPK students may only have perceived options as attractive if they had good reasons for it (knowledge as main source), and thus only when they seriously considered them as being correct, explaining why these options received much attention. LPK students, in contrast, considered more answer options as being attractive without having good reasons for it (intuition as main source), thereby being potentially more prone to dismissing them without deliberate consideration, explaining why these options received less attention. Further research is needed to provide more direct evidence in favor of these claims, and thus to explain the obtained interaction concerning the gaze bias hypothesis satisfactorily.

According to the *gaze bias consolidation hypothesis*, we expected the gaze bias to occur especially in the final phase of the decision process when solving MC items, and thus, right before the decision announcement. Moreover, this gaze bias was expected to occur in a similar manner for students with different levels of domain knowledge. As expected, results of our time course analysis showed that between the last three time intervals of the item solution process, HPK and LPK students' fixation times increased only for chosen options, while fixation times for attractive and non-attractive options (i.e., not chosen options) remained stable or even decreased. Thus, these results are again in line with prior research on eye movements and decision making, yielding support for the finding that a gaze bias occurs primarily in the final stage of the decision process (cf., Shimojo et al., 2003).

However, even in the literature on basic decision research it is yet an open question which specific cognitive processes are reflected by the gaze bias effect (e.g., Glaholt & Reingold, 2011; Schotter et al., 2010; Shimojo et al., 2003). Nonetheless, our data may contribute to the rejection of the assumption that the gaze bias is just a response-related phenomenon (e.g., Glaholt & Reingold, 2009b), because the gaze bias started to occur quite early in the present study. Moreover, students showed a tendency to fixate more on chosen than on non-chosen options during the entire decision process, and thus even before the strong gaze bias occurs in the last three time intervals prior to the choice announcement. This suggests that the gaze bias did not solely reflect the motor response programming that *follows* the cognitive process of decision making, but rather, that it may reflect a cognitive state where the decision has already been made on an unconscious level, while choice awareness may follow from the gaze bias. Another explanation might be that the choice has already consciously been made and the gaze bias effect then reflects a post-decision evaluation of the chosen answer option. However, while these different explanations are intriguing and worthy of further investigation, they are still speculative and lack empirical validation.

In line with our assumption that the gaze bias consolidation would occur in a comparable manner for students with different domain knowledge levels, results of the time course analysis revealed that the time course of fixation time on non-attractive, attractive, and chosen options followed similar patterns for both HPK and LPK students (cf., Figure 4 and 5). The slopes of the fixation time curves only differed in the beginning or at the end of the time course (in only 4 out of 36 time intervals). Thus, the seemingly high comparability of the curves for HPK and LPK students in the diagrams (cf., Figures 4 and 5) is furthermore quantitatively supported by the results of the repeated contrasts.

At first sight, results of similar curve patterns across the solution time for HPK and LPK students may seem to contradict the interaction between *preference* and *domain knowledge* (cf., results of the *gaze bias hypothesis*). These findings, however, can be reconciled by arguing that even though the overall amount of attention on answer options with distinct preference ratings may differ, the amount of attention on these options as a function of processing time may nevertheless be similar. Another note concerns the division of the total time on task into ten equal time bins that cover the whole solution process. This method covers absolute differences in real solution time between HPK and LPK students. Thus, it needs to be taken into consideration that the actual time span of the decision process was overall shorter for HPK than for LPK students (cf., reaction time comparison).

Nevertheless, the similar qualitative pattern of attention allocation over time provides tentative evidence in favor of the claim that both groups of students went through similar cognitive stages in the item-solving process, even though their duration might be longer or shorter in relation to real time. These cognitive stages may relate to four distinct decision making stages formulated within the model of Russo and Leclerc (1994), which originates from research on consumers' decision making. According to Russo and Leclerc (1994), the four stages are termed (1) *screening and orienting*, (2) *deliberate evaluation* (3a) *review and choice announcement* and (3b) a *post-announcement review stage* (time after the decision has been announced, which cannot be observed in our data because this time was eliminated from analysis). With regard to the decision making process in the MC test of the present study, students mostly attended to the question and performed a first screening of available answer options in the first three time intervals, thus reflecting a *screening and orienting* phase. In the next four time intervals students paid less attention to the question and focused more on the answer alternatives, possibly reflecting a *deliberate evaluation* of all four answer options regarding their correctness. In the last three time intervals prior to the *choice announcement*, the gaze bias towards the chosen options occurred in the present study which possibly also includes a *review* of students' actual choice.

To conclude, in line with our hypotheses, results indicate that the decision making process, which underlies MC item-solving, is largely comparable for HPK and LPK students when referring to subjective preference. By contrast, prior eye tracking studies on the processing of MC items found differences in item-solving behavior depending on the knowledge levels of students (e.g., Andrà et al., 2009; Holmqvist, Andrà et al., 2011; Tai et al., 2006; Tang & Pienta, 2012; Tsai et al., 2012). This apparent conflict might go back to the fact that prior studies focused primarily on *objective* criteria of item-solving, while furthermore applying different analysis methods. The present study, however, was the first to collect and also take into account data of students' *subjective* preferences for answer options.

Thus, by focusing on two different analysis levels, the present data revealed a *knowledge identification* function of eye fixations when related to the (objective) correctness of answer options as well as a *gaze bias effect* that was strongly related to students' (subjective) preferences for answer options. Furthermore, prior studies used MC material with a stronger focus on problem solving (e.g., mathematics) that often comprised graphic elements like formulas, diagrams or pictures. However, in this study, we used verbal, knowledge assessing MC items. Taken together, the differences in both the focus of the research and the item characteristics probably led to the apparent differences in results and conclusions between prior studies on this topic and the present research.

5.1 Limitations and Implications for Future Research

As a potential limitation of this study, one might notice the sample size of twenty-six students. Even though this sample size is moderate on a person level, it is to acknowledge that all analyses are based on an item level as a within-subject factor, resulting in a total number of 546 analysis units (26 students multiplied by 21 test items). Especially in combination with the use of an eye tracker that records high-frequency, objective data of students' processing behavior, in the present study the analyses relied on a sufficiently large data base to produce reliable effects. Therefore, we would not expect an entirely different pattern of results with a larger sample size on a person level.

A first implication of our study is that the results of the present research may be relevant not only for the field of *CDA*, but also for *decision making research*. By analyzing the eye movement data, we obtained detailed insights into how students processed the MC test material and we were able to partially explain students' solution behavior by referring to the literature on knowledge-related cognitive processing and decision making.

Concerning CDA, the present data suggest that eye movements can indeed help to uncover item-solving processes and may thus help to detect interactions between student characteristics and test characteristics (cf., Gorin, 2007). More precisely, our results provide basic information about how students with different knowledge levels process the different components of MC items (question, correct and incorrect answer options) over time and how their preferences for answer options interplay with the attention devoted to them. Thus, our data contribute towards an understanding of MC item processing on a fundamental theoretical level and might thereby provide a starting point for future eye tracking research in the field of CDA.

Concerning the area of decision making research, in the present study, both HPK and LPK students showed a similar gaze bias towards preferred and chosen options in a knowledge-assessing MC test. Thus, the gaze bias was found to be a stable phenomenon across different levels of prior knowledge in MC testing, thereby providing further evidence for the generalizability of the gaze bias effect across various decision tasks. Nevertheless, it is important to replicate the present findings in different settings and with different MC tests (e.g., different content or item formats), as well as with different samples of participants. For example, on a theoretical level, it would be interesting to examine how *experts* (e.g., professors or PhDs with outstanding domain knowledge; cf., Jarodzka et al., 2010) process an MC test in their expertise domain. One could hypothesize that the MC item-solving process of experts might be similar to that of HPK students in the present study, except that the gaze bias and attention focusing effects on correct answer options might be much more pronounced.

A second, more applied implication of the current results is that eye movements have been shown to have a high potential to predict choice behavior and attitudes towards answer options in MC items. Specifically, we were able to identify HPK students' *knowledge levels* by analyzing how much attention they devoted to correct answer options, especially in the later phase of the decision process. This is interesting because gaze allocation and duration may provide a new source of information, allowing for a better distinction of students with higher and lower domain knowledge levels in future educational practice. Even though such applications might seem quite unrealistic at the moment, the use of eye tracking technology in classrooms could become a reality sooner or later (e.g., digital classroom project in Lund, Sweden; cf., ‘The Digital classroom: A new world-class lab’, 2012).

Furthermore, combined with similar efforts and results from Bee et al. (2006) or Glaholt et al. (2009), the present data provide tentative evidence in favor of using eye movements to evaluate students’ *preferences* by analyzing their gaze duration on answer options. This approach, especially in combination with retrospective think aloud protocols (cf., Van Gog, Paas, Van Merriënboer, & Witte, 2005), could, for example, help to detect item flaws (e.g., improper wording or misconceptions) that cannot be taken from test scores alone. In the future, such an application might be employed to support distractor analyses or to investigate test validity, especially in high-stakes and large-scale assessments. Thus, eye tracking might become a sophisticated tool in the construction of decisive assessments such as PISA (OECD, 2013), and might contribute towards the development of high-quality test items.

However, before promising practical implementations like these can really be implemented in educational settings, further research is certainly required to check whether eye movement recordings in the context of testing can actually live up to their potential.

References

Amadieu, F., Van Gog, T., Paas, F., Tricot, A., & Mariné, C. (2009). Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning. *Learning and Instruction*, *19*, 376–386. doi:10.1016/j.learninstruc.2009.02.005

Andrà, C., Arzarello, F., Ferrara, F., Holmqvist, K., Lindstrom, P., Robutti, O., & Sabena, C. (2009). How students read mathematical representations: An eye tracking study. In M. Tzekaki, M. Kaldrimidou, & H. Kaldrimidou (Eds.), *Proceeding of the 33rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 49–56). Thessaloniki, Greece: Organising Commitees.

Bee, N., Prendinger, H., Nakasone, A., André, E., & Ishizuka, M. (2006). AutoSelect: What you want is what you get: Real-time processing of visual attention and affect. In E. André, L. Dybkjær, W. Minker, H. Neumann, & M. Weber (Eds.), *Perception and Interactive Technologies* (pp. 40–52). Berlin: Springer. doi:10.1007/11768029_5

Bird, G. D., Lauwereyns, J., & Crawford, M. T. (2012). The role of eye movements in decision making and the prospect of exposure effects. *Vision Research*, *60*, 16–21. doi:10.1016/j.visres.2012.02.014

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain.* New York, NY: David McKay.

Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction*, *20*, 155–166. doi:10.1016/j.learninstruc.2009.02.014

Duchowski, A. (2007). *Eye tracking methodology: Theory and practice*. London, UK: Springer.

Ellis, J. J., Glaholt, M. G., & Reingold, E. M. (2011). Eye movements reveal solution knowledge prior to insight. *Consciousness and Cognition*, *20*, 768–776. doi:10.1016/j.concog.2010.12.007

Embretson, S. E. (1999). Cognitive psychology applied to testing. In F. T. Durso (Ed.), *Handbook of Applied Cognition* (pp. 629–660). Chinchester, UK: Wiley.

Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343–368. doi:10.1111/j.1745-3984.2001.tb01131.x

Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*, 1035–1045. doi:10.1016/S0042-6989(03)00084-1

Flowe, H. (2011). An exploration of visual behaviour in eyewitness identification tests. *Applied Cognitive Psychology*, *25*, 244–254. doi:10.1002/acp.1670

Flowe, H., & Cottrell, G. W. (2011). An examination of simultaneous lineup identification decision processes using eye tracking. *Applied Cognitive Psychology*, *25*, 443–451. doi:10.1002/acp.1711

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, *23*, 523–552. doi:10.1007/s10648-011-9174-7

Glaholt, M. G., & Reingold, E. M. (2009a). Stimulus exposure and gaze bias: A further test of the gaze cascade model. *Attention, Perception, & Psychophysics*, *71*, 445–450. doi:10.3758/APP.71.3.445

Glaholt, M. G., & Reingold, E. M. (2009b). The time course of gaze bias in visual decision tasks. *Visual Cognition*, *17*, 1228–1243. doi:10.1080/13506280802362962

Glaholt, M. G., & Reingold, E. M. (2011). Eye movement monitoring as a process tracing methodology in decision making research. *Journal of Neuroscience, Psychology, and Economics*, *4*, 125–146. doi:10.1037/a0020692

Glaholt, M. G., Wu, M.-C., & Reingold, E. M. (2009). Predicting preference from fixations. *PsychNology Journal*, *7*, 141–158.

Glaholt, M. G., Wu, M.-C., & Reingold, E. M. (2010). Evidence for top-down control of eye movements during visual decision making. *Journal of Vision*, *10*(5), 1–10. doi:10.1167/10.5.15

Gorin, J. S. (2007). Test construction and diagnostic testing. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive Diagnostic Assessment in Education: Theory and Practice.* Cambridge, United Kingdom: Cambridge University Press.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. New York, NY: Routledge.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*, 309–334. doi:10.1207/S15324818AME1503_5

Healy, A. F. (Ed.). (2005). *Experimental cognitive psychology and its applications*. *Decade of behavior.* Washington, DC: American Psychological Association. doi:10.1037/10895-000

Hegarty, M., Canham, M. S., & Fabrikant, S. I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 37–53. doi:10.1037/a0017683

Heller, K. A., & Perleth, C. (2000). *KFT 4-12+R - Kognitiver Fähigkeits-Test für 4. bis 12. Klassen, Revision*. Göttingen, Germany: Beltz.

Holm, S., (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70. Retrieved from http://www.jstor.org/stable/4615733

Holmqvist, K., Andrà, C., Lindström, P., Arzarello, F., Ferrara, F., Robutti, O., & Sabena, C. (2011). A method for quantifying focused versus overview behavior in AOI sequences. *Behavior Research Methods*, *43*, 987–998. doi: 10.3758/s13428-011-0104-x

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & De Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford, United Kingdom: Oxford University Press.

Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction*, *20*, 172–176. doi:10.1016/j.learninstruc.2009.02.013

Jarodzka, H., Scheiter, K., Gerjets, P., & Van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, *20*, 146–154. doi:10.1016/j.learninstruc.2009.02.019

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329–355. doi:10.1037/0033-295X.87.4.329

Kaakinen, J. K., Hyönä, J., & Keenan, J. M. (2003). How prior knowledge, WMC, and relevance of information affect eye fixations in expository text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 447–457. doi:10.1037/0278-7393.29.3.447

Kim, S., & Rehder, B. (2011). How prior knowledge affects selective attention during category learning: An eyetracking study. *Memory & Cognition*, *39*, 649-665. doi:10.3758/s13421-010-0050-3

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857. doi:10.1073/pnas.1101328108

Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature*, *384*, 74–77. doi:10.1038/384074a0

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15. doi:10.1111/j.1745-3992.2004.tb00164.x

Leighton, J. P., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education. Theory and applications*. Cambridge, United Kingdom: Cambridge University Press.

Lohse G. L., & Johnson, E. J. (1996). A comparison of two process tracing methods for choice tasks. *Organizational Behavior and Human Decision Processes*, *68*, 28–43. doi.10.1006/obhd.1996.0087

Mansour, J. K., Lindsay, R. C. L., Brewer, N., & Munhall, K. G. (2009). Characterizing visual behaviour in a lineup task. *Applied Cognitive Psychology*, *23*, 1012–1026. doi:10.1002/acp.1570

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–

103). New York, NY: Macmillan.

Mitsuda, T., & Glaholt, M. G. (2014). Gaze bias during visual preference judgements: Effects of stimulus category and decision instructions. *Visual Cognition, 22*(1), 11–29. doi:10.1080/13506285.2014.881447

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, *64*, 575–603. doi:10.3102/00346543064004575

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. New York, NY: Routledge.

OECD (2013). *PISA 2012 assessment and analytical framework. Mathematics, reading, science, problem solving and financial literacy*. Paris. OECD. doi:10.1787/9789264190511-en

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press. doi:10.17226/10019

Pieters, R., & Warlop, L. (1999). Visual attention during brand choice: The impact of time pressure and task motivation. *International Journal of Research in Marketing*, *16*, 1–16. doi:10.1016/S0167-8116(98)00022-6

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*, 3–25. doi:10.1080/00335558008248231

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*, 160–174. doi:10.1037/0096-3445.109.2.160

Reutskaja, E., Nagel, R., Camerer, C. F., & Rangel, A. (2011). Search dynamics in consumer choice under time pressure: An eye-tracking study. *The American Economic Review*, *101*, 900–926. doi:10.1257/aer.101.2.900

Russo, J. E. (1978). Eye fixations can save the world: A critical evaluation and a comparison between eye fixations and other information processing methodologies. *Advances in Consumer Research*, *5*, 561–570.

Russo, J. E., & Leclerc, F. (1994). An eye-fixation analysis of choice processes for consumer nondurables. *Journal of Consumer Research*, *21*, 274–290.

Saß, S., Wittwer, J., Senkbeil, M., & Köller, O. (2012). Pictures in test items: Effects on response time and response correctness. *Applied Cognitive Psychology*, *26*, 70–81. doi: 10.1002/acp.1798

Schotter, E. R., Berry, R. W., McKenzie, C. R. M., & Rayner, K. (2010). Gaze bias: Selective encoding and liking effects. *Visual Cognition*, *18*, 1113–1132. doi:10.1080/13506281003668900

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, *6*, 1317–1322. doi:10.1038/nn1150

Simion, C., & Shimojo, S. (2006). Early interactions between orienting, visual sampling and decision making in facial preference. *Vision Research*, *46*, 3331–3335. doi:10.1016/j.visres.2006.04.019

Sweller, J., Van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251–296. doi:10.1023/A:1022193728205

Tai, R. H., Loehr, J. F., & Brigham, F. J. (2006). An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. *International Journal of Research and Method in Education*, *29*, 185–208. doi:10.1080/17437270600891614

Tang, H., & Pienta, N. (2012). Eye-tracking study of complexity in Gas Law Problems. *Journal of Chemical Education*, *89*, 988–994. doi:10.1021/ed200644k

The Digital classroom: A new world-class lab (2012, November 15). Retrieved from http://www.lunduniversity.lu.se/o.o.i.s?news_item=5954&id=24890

Thoma, G.-B., Dalehefte, I. M., & Köller, O. (2014). Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung von Wissen über das menschliche Gehirn und Nervensystem. *Psychologie in Erziehung und Unterricht, 61*, 231-236. doi:10.2378/peu2014.art18d

Tsai, M.-J., Hou, H.-T., Lai, M.-L., Liu, W.-Y., & Yang, F.-Y. (2012). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Computers & Education*, *58*, 375–385. doi:10.1016/j.compedu.2011.07.012

Van Gog, T., Paas, F., & Van Merriënboer, J. J. (2005). Uncovering expertise-related differences in troubleshooting performance: Combining eye movement and concurrent verbal protocol data. *Applied Cognitive Psychology*, *19*, 205–221. doi:10.1002/acp.1112

Van Gog, T., Paas, F., Van Merriënboer, J. J., & Witte, P. (2005). Uncovering the problem-solving process: cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied*, *11*, 237–244. doi:10.1037/1076-898X.11.4.237

Wright, R. D., & Ward, L. M. (2008). *Orienting of attention*. New York, NY: Oxford University Press.