Advances in Health Sciences Education (2021) 26:1339–1354
https://doi.org/10.1007/s10459-021-10052-z

Do different response formats affect how test takers approach a clinical reasoning task? An experimental study on antecedents of diagnostic accuracy using a constructed response and a selected response format

Stefan K. Schauber[1] · Stefanie C. Hautz[2] · Juliane E. Kämmer[3,4] · Fabian Stroben[4,5] · Wolf E. Hautz[2,6]

Received: 26 September 2018 / Accepted: 3 May 2021 / Published online: 11 May 2021

Abstract
The use of response formats in assessments of medical knowledge and clinical reasoning continues to be the focus of both research and debate. In this article, we report on an experimental study in which we address the question of how much list-type selected response formats and short-essay type constructed response formats are related to differences in how test takers approach clinical reasoning tasks. The design of this study was informed by a framework developed within cognitive psychology which stresses the importance of the interplay between two components of reasoning—self-monitoring and response inhibition—while solving a task or case. The results presented support the argument that different response formats are related to different processing behavior. Importantly, the pattern of how different factors are related to a correct response in both situations seem to be well in line with contemporary accounts of reasoning. Consequently, we argue that when designing assessments of clinical reasoning, it is crucial to tap into the different facets of this complex and important medical process.

Keywords Clinical reasoning · Response format · Cognitive reflection · Processing fluency

✉ Stefan K. Schauber
stefan.schauber@medisin.uio.no

[1] Centre for Health Sciences Education, Faculty of Medicine, University of Oslo, Postboks 1161 Blindern, 0318 Oslo, Norway

[2] Department of Emergency Medicine, Inselspital University Hospital, University of Berne, 3010 Freiburgstrasse, Berne, Switzerland

[3] Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

[4] AG Progress Test Medizin, Charité Medical School Berlin, Hannoversche Straße 19, 10115 Berlin, Germany

[5] Office of the Vice Dean for Teaching and Learning, Charité Universitätsmedizin, Berlin, Germany

[6] Centre for Educational Measurement (CEMO), University of Oslo, Postboks 1161 Blindern, 0318 Oslo, Norway

Introduction

A crucial decision when designing an educational assessment is the selection of the format in which test takers give their response to the tasks or problems posed. Both research on assessment in medical education (Desjardins et al., 2014; Heemskerk et al., 2008; Huwendiek et al., 2017; Norman et al., 1996; Sam et al., 2018; Schuwirth et al., 1996) and the broader literature on educational and psychological testing (Bleske-Rechek et al., 2007; DeMars, 1998, 2000; Hickson et al., 2012; Lukhele et al., 1994; Rodriguez, 2003) have repeatedly focused on the question of in how far different item response formats evoke differences in task processing behavior.

Scholars within the field of educational testing typically draw a distinction between constructed and selected response formats. Selected response (SR) formats include all types of questions where test takers have to pick the correct option(s) out of a list. These are, for instance, multiple-choice, true/false, or multiple response questions. Constructed response (CR) formats, on the other hand, are questions or tasks where test takers have to generate the answer on their own. For example, this is the case in any essay or simply when test takers have to write in a single term (e.g. a diagnosis). Indeed, there is a common suspicion that these two broader classes of response formats evoke fundamentally different cognitive processes. Recognizing the correct option is perceived as being very different from generating the answer (Lissitz et al., 2012; Martinez, 1999; Ozuru et al., 2013). The main concern here is that the use of selected response formats would compromise the validity of assessments. If true, this would have important implications for both the design of assessments in medical education and research on clinical reasoning. However, empirical data to support these claims are limited (Hift, 2014; Norman et al., 1996).

On a pragmatic level, research shows that variations in response formats hardly affect actual assessment outcomes. That is, across disciplines, studies typically find high correlations between performances on tests using SR and CR formats (Martinez, 1999; Rodriguez, 2003). For example, in a recent article, Desjardins et al. (2014), found a correlation of $r = 0.83$ between performances on SR and CR tests using identical item-stems (i.e., identical clinical cases) in both conditions. At the same time, studies also report a consistent difference—selected response items are typically easier to answer. Compared to constructed response formats, test takers answer correctly more often when a SR format is used even if the problem posed is identical (Norman et al., 1996; Sam et al., 2018; Schuwirth et al., 1996). However, research also shows that using a SR format can make a task more difficult. This occurs when the options contain a highly attractive, but incorrect answer (Desjardins et al. 2014; Schuwirth et al., 1996). Importantly, these findings clearly highlight that test takers indeed make use of the options presented in a selected response format—even though the same options might, at times, be misleading.

Obviously, the possibility to uncover differences and similarities necessarily depends on the framework used to conceptualize response behavior or processes. Typically, studies conducted within medical education have been based on dual-process models. For instance, one popular framework differentiates between inductive and deductive reasoning (Elstein et al., 1978; Patel et al., 1993). Hence, the according studies use think-aloud techniques to reveal this type of reasoning (Heemskerk et al. 2008). Others focus on the reasoning as either intuitive/fast or elaborate/slow (Monteiro & Norman, 2013); in the according studies, analysis of response times plays a crucial role (Monteiro et al. 2015). In this paper, we aim at offering a new perspective on response behavior in a clinical reasoning scenario, which, in turn, introduces a new methodological approach, too.

In this study, we use a more recent framework that focusses on understanding the origins of errors in reasoning (De Neys, 2013, 2014; De Neys & Bonnefon, 2013; De Neys & Glumicic, 2008). DeNeys' approach is rooted in dual process theory as it assumes that a response can be rather intuitive or more elaborate. The authors postulate that errors can originate from three elementary components of reasoning: storage, self-monitoring, and response inhibition. The first component, storage, means that reasoners answer incorrectly because they simply might not know or know wrong. Incorrect knowledge or misconceptions are, however, typically acquired long before an actual task is processed. The second crucial component is self-monitoring, which occurs while working on a problem. For instance, the feeling of being confident can determine the course of how a participant engages in solving a task or problem (Thompson et al., 2011). Critically, the third and final component is the ability to inhibit an intuitive response. Such inhibition is regarded as a key element in order to be able to adapt the actual reasoning process, that is, to switch from an intuitive response to a more elaborate one. While this framework resembles current thinking in research on clinical reasoning (e.g., Norman & Eva, 2010), there are a number of critical differences and additions, especially in regard to how self-monitoring and inhibition are understood.

The first difference is related to how various measures self-monitoring are subsumed in one single indicator of task-fluency. Task fluency is a reasoning person's experience of processing a problem (Benjamin et al., 1998; Oppenheimer, 2008). This perception is formed by three indicators: The appraisal of something being difficult, which leads to increased time-on-task, which then results in a judgment of low confidence (Alter & Oppenheimer, 2009; Dunlosky, & Thiede, 2013; Hertwig et al., 2008; Koriat, 2012; Kornell et al., 2011). These three aspects, taken together, are indicators of task fluency, which, in turn, is a crucial trigger that can alter the reasoning process. For instance, reasoners might engage in more elaborate reasoning if they perceive low task fluency (Alter et al., 2007). This idea has its counterpart in research on overconfidence in diagnostic reasoning. Here too, the ability to appraise one's own lack of knowledge or expertise is assumed to be crucial for avoiding—sometimes serious—errors (Berner & Graber, 2008). In summary, while there are many similarities to the concept of self-monitoring, task fluency allows for integrating different measures into a theoretically informed single indicator.

Importantly, Second, the approach by DeNeys and colleagues assumes that inhibiting an initially attractive, intuitive response is a crucial capacity in any reasoning scenario (Pennycook et al., 2015; Thompson et al., 2013; Toplak et al., 2014). Much of the thinking on such inhibitory processes is related to a paradigm used in the Cognitive Reflection Test (Fredrick, 2005). The starting point in these studies is to create tasks or problems that tend to trigger an immediate, but incorrect response. In order to solve a problem correctly, participants have to inhibit and overwrite this spontaneous response. For instance, one of the tasks in the cognitive reflection test goes as follows: *A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?* The prepotent, incorrect, response in this case is "10 cents", and, indeed, many participants give this intuitive response (Toplak et al., 2011). In order to answer correctly to such questions (assuming that they do not already know the solution), participants have to realize that the intuitive solution is incorrect, and overwrite their first response.

Since its first publication, the Cognitive Reflection Test (CRT) has been used in a wide range of studies, and substantial correlations to a number of indicators of faulty or biased reasoning have been found (Białek & Sawicki, 2018). In its original conception, the validity of the claim that this test indeed is an indicator of the ability to inhibit spontaneous responses has been mostly justified on a theoretical basis by the design of the tasks or questions. More recently, studies have focused on tracing participants' response behavior when working through problems where they have to inhibit a prepotent answer (Toplak et al., 2011; Travers et al., 2016).

Indeed, the ability to inhibit and reflect on a prepotent answer may play a crucial, but differential, role in any testing scenario. In the particular context discussed here, SR and CR formats presumably differ with respect to the demands they set for test takes. Two strands of research support this expectation. First, previous studies found that more intuitive response behavior is more likely observed in SR format conditions (Heemskerk et al., 2008). Second, studies by Kostopoulou et al. (2012, 2015, 2016) found that reconsidering an initial hypothesis improves accuracy. In her studies, presenting a list of likely diagnoses was an especially effective intervention. Furthermore, Mamede and colleagues repeatedly demonstrated that deliberately reflection on difficult cases improves diagnostic accuracy, too (Mamede et al, 2012, 2020). Taken together, these findings suggest that, firstly, responses in SR conditions are more likely to be given intuitively but also that, secondly, reconsidering a spontaneous answer can improve accuracy.

The research reviewed above highlights an obvious contradiction. On the one hand, studies largely find that the scores obtained from assessments using one of the two formats are highly correlated—suggesting that the formats are largely exchangeable. At the same, both theoretical approaches and empirical evidence support the stance that different response formats trigger differences in cognitive demands and response behavior. Indeed, very different processes might still lead to the same scores.

Against this backdrop, we aim at addressing the question of why scores differ between CR and SR formats from a new perspective. We use DeNeys' account of three elementary factors of reasoning to formulate expectations on where in the response process such differences actually occur. This account introduces a new perspective on investigating the effects of response formats in clinical reasoning tasks. It critically adds in two distinct ways theoretically and methodologically. First, it opens up the perspective that different response formats can be understood as setting different demands to the test taker, but, crucially, neither of them is inherently "more" or "less" valid. Second, this perspective builds on the conception of fast/slow thinking, but also goes beyond it in an important way. By stressing the importance of inhibition and, consecutively, reflection in intuitive response behavior, this framework accounts for some of the criticism dual process models are faced with (Evans, 2008; Kruglanski & Gigerenzer, 2011). Hence, the current study adds to the literature in two critical ways:

- On a theoretical level, we introduce a contemporary framework from cognitive psychology to health science education in order to reframe the issue on differences in response formats.
- On an empirical level, we formulate and investigate research questions on the role of these components in processing clinical reasoning tasks.

We conducted an experimental study in which we investigate the effect of a selected response format versus a constructed response format on response behavior in a clinical reasoning scenario. In our study, we assume that case-specific knowledge was independent of the testing condition, as we conducted a randomized study. Furthermore, we address two main research questions. First, we expect that differences in perceived fluency are related to differences in accuracy across cases. Second, we expect that scores on the CRT—as an indicator of the ability to inhibit a response—are differently related to accuracy across the two experimental conditions (CR, SR). Since one of the concerns frequently raised in regard to SR questions is that testees could simply guess the correct answer, we formulate a fourth research question. We speculate that reported guessing is related to low perceived fluency and consequently to lower chances of success.

Methods

Participants

The study was conducted at Charité–Universitätsmedizin Berlin. We invited all 350 medical students in their 4th academic year via email and through Facebook postings to participate in a "study of factors affecting decision making in emergency medicine". Participation was voluntary and the institutional review board of Charité—Universitätsmedizin granted the study its approval (EA4/096/16). The first 60 replying students were invited to participate in the study. A total number of N = 54 students (67% females) took part. On average, participants were M = 24.6 (SD = 3.38) years old.

Procedure

Upon arrival, participants were randomly assigned to one of two experimental conditions – CR (N = 27) or SR (N = 27). Participants were then informed about the study procedure and signed the consent form. After filling in a questionnaire on general demographic information, participants received a demonstration of how to work on the clinical cases. In total, one training case plus six clinical cases were administered, the latter were presented in random order. After participants had completed the cases, the Cognitive Reflection Test (Frederick, 2005) was administered. The complete session lasted about one hour, for which participants were compensated with €20 ($22 at that time).

Clinical cases

We administered the ASCLIRE (Kunina-Habenicht et al., 2015) assessment, which consists of six clinical cases plus one trial case. Each case presents a patient with shortness of breath. All cases depict common causes of acute and sub-acute dyspnea. Participants were instructed to take all diagnostic tests they deemed relevant but no more than that. A total number of 30 diagnostic tests are available to choose from. Clicking on a test elicited the finding in the form of a text (e.g., pulse rate), an image (e.g., ECG, chest X-ray), or audio (e.g., heart sounds, history). Where feasible, these findings require the participants' interpretation (e.g., ECG, heart sounds). Some findings (e.g., ultrasound exams, CT scans) are available only as radiologists' textual report. Participants were free to choose any type, order and number of diagnostic tests they wanted to see or listen to, and repeated acquisition was allowed.

Participants were instructed to diagnose the patient as fast as possible without sacrificing accuracy. Importantly, students could decide to end the information-gathering phase and move on to giving their diagnosis. Once they decided to move on, they were no longer able to obtain clinical tests for the case. This procedure allowed for separating the time students took for obtaining information and processing the case form the time they needed to enter their diagnosis. After submitting their diagnosis and before proceeding to the next case, all participants were asked to evaluate each case with regard to its difficulty, whether or not they were guessing, and to what extent they were confident in the correctness of the diagnosis.

In the selected response condition, participants were free to choose a diagnosis by selecting one out of a list of 20. No return to the diagnostic tests was possible at this point. The list of possible diagnoses was the same for all cases and ordered alphabetically. In the constructed response condition, participants entered their diagnosis into a free text form after they finished processing the case. No return to the diagnostic tests was possible at this point. Three board certified emergency physicians with each at least 10 years of professional experience independently evaluated each CR response, blinded towards which examinee provided it. The final accuracy score was then derived from the majority of raters.

In the study by Kunina-Habenicht et al. (2015), a Cronbach's Alpha of Alpha = 0.48 across the six cases was reported. The authors provide further evidence for the validity of the ASCLIRE framework in the said article.

Measures

Accuracy

The main outcome measure was whether or not students found the correct diagnoses to the presented cases. Accuracy of the selected diagnoses was treated as a dichotomous measure (correct or incorrect).

Conflict detection and indicators of task fluency

The meta-cognitive measures obtained were confidence and perceived case difficulty. Furthermore, the time on a case was recorded in seconds. Confidence in the correctness of the diagnosis was rated on a percent-scale from 0% (*no confidence*) to 100% (*highest confidence)* in 10% increments. Perceived case difficulty was evaluated on a 5-point rating scale from 1 (*very easy*) to 5 (*very difficult*).

Combined fluency score

In order to build a combined fluency score, self-reports on confidence and difficulty as well as time spent on case were standardized within each participant across cases with a mean of $M_{within} = 0$ and a standard deviation of $SD_{within} = 1$ (i.e.,within-person centred). We then multiplicated the z-standardized time-on-case and difficulty ratings by minus 1, thus reversing these two measures. In this way, the three variables had the same interpretation with regard to the fluency measure: Higher values on the three variables signified higher fluency. After centering and reversing, an average score was calculated using the three variables within every case. As a result, we obtained one fluency score for each person on each case (where higher scores indicate higher fluency). Hence, these scores carry information on the relative fluency experienced between cases and within each person.

Cognitive reflection test

We administered a German version of the three items cognitive reflection test (Frederick, 2005) after students completed the six clinical cases. The score on the CRT was calculated as the number of correct responses on the three-item test. Reliability was determined by means of Cronbach's Alpha.

Guessing

Participants reported guessing per case dichotomously. After submitting a diagnosis, they received the following prompt: "Thank you for your diagnosis. Did you guess it? Yes/no".

Analytic procedure

Generalized linear mixed models (GLMMs) were used to analyze between-group differences in chances to solve a case correctly. The models used had the general form of:

$$logit\left(P_{ij}\right) = \gamma_0 + \sum_{h=1}^{r} \gamma_h x_{hij} + S_{0i} + C_{0j}$$

where P_{ij} is the odds ratio for subject i giving a correct response to case j,γ_0 indicates the intercept, and S_{0i} and C_{0j} represent the random intercepts for subjects and cases, both following a normal distribution with a mean of 0 and standard deviations of τ_{00} and ω_{00}, respectively. As usual, the residual term in a logistic model is fixed to $\frac{\pi^2}{3}$ (≈3.29) and hence remains constant across all models. The sum $\sum_{h=1}^{r} \gamma_h x_{hij}$ represents the $X_{(h=1,\ldots,r)}$ predictors while γ_h represents the according fixed effect. This means that x_{hij} represents the value of subject i on case j for predictor X_h. For example, $x_{4,3,2}$ would signify subject 3´s response on case 2 to the question of whether she was guessing the diagnosis on that case or not ("0" or "1"). The $r=4$ predictors (response format, CRT, task fluency, and reported guessing) were entered successively meaning that, in total, 4 increasingly complex models were estimated.

We calculated the explained variance at the level of the random effects by calculating the proportional reduction of variance at the given level in relation to a Null Model (i.e., a model including an intercept only). Details on this procedure can be found in Snijders and Bosker (2011). The threshold for statistical significance was set at $p=5\%$. The package lme4 (Bates et al., 2015) within the R Language and Environment for Statistical Computing (R Core Team, 2018) was used to estimate the models.

Results

Descriptive statistics

Please refer to Table 1 for descriptive statistics for both groups and all cases. Case 4 ('pneumonia') was the easiest case, diagnosed correctly by 93% of the participants in the

Table 1 Descriptive statistics for the six presentations per group

Clinical case	Condition	Diagnostic accuracy	Reported Guessing	Confidence	Experienced difficulty	Time on case (in seconds)	Time for submitting response (in seconds)	Fluency (z-scores)
1: instable ventricular tachycardia	SR	44%	26%	62.96 (18.77)	3.48 (0.94)	158.95 (82.98)	26.64 (15.69)	−0.06 (0.65)
	CR	23%	15%	60.38 (23.91)	3.58 (0.95)	160.22 (94.22)	16.83 (26.04)	
2: chronic obstructive pulmonary disease	SR	93%	15%	76.67 (16.64)	2.52 (1.12)	125.09 (65.5)	9.97 (10.42)	0.34 (0.68)
	CR	77%	4%	66.54 (19.58)	3.23 (0.82)	171.06 (78.86)	21.35 (12.34)	
3: pulmonary edema	SR	52%	15%	64.44 (25.47)	3.41 (1.01)	168.01 (76.24)	29.77 (31.1)	−0.23 (0.58)
	CR	31%	12%	54.23 (22.48)	3.88 (0.65)	249.53 (144.85)	17.88 (17.6)	
4: pneumonia	SR	93%	11%	78.15 (15.94)	2.59 (1.05)	124.32 (66.14)	15.62 (11.52)	0.41 (0.52)
	CR	92%	15%	66.15 (20.8)	2.85 (0.83)	156.6 (98.5)	13.98 (6.11)	
5: pulmonary artery embolism	SR	59%	22%	62.96 (27.29)	3.56 (1.22)	176.05 (105.64)	22.15 (23.8)	−0.06 (0.71)
	CR	50%	15%	60.38 (24.41)	3.62 (0.98)	174.48 (82.1)	13.87 (15.48)	
6: intoxication	SR	56%	30%	57.04 (25.99)	3.7 (0.95)	176.74 (124.57)	23.45 (20.3)	−0.39 (0.78)
	CR	42%	31%	46.92 (26.04)	3.96 (1.08)	214.14 (112.3)	25.92 (30.78)	

Mean for diagnostic accuracy, reported guessing (both in percent). Mean (SD) for confidence, experienced difficulty and time on case. Mean (SD) for the combined fluency scores on the case-level where higher scores indicate higher fluency

SR group and by 92% in the CR group. Across the two conditions, the average Pearson correlation of accuracy across all six cases across the experimental conditions was 0.09, and ranged between a minimum of –0.05 (case two and case six) and a maximum of 0.28 between case 3 (pulmonary edema) and case 6 (intoxication). For the Cognitive Reflection Test, Cronbach's Alpha was $\alpha = 0.81$.

Table 1 gives descriptive statistics for the key measures in this study. For instance, the highest observed difference in guessing between groups were found for Case 2 ('COPD') and Case 1 ('instable ventricular tachycardia'). For Case 2, 15% of the participants in the SR group (92% correct) reported guessing as opposed to 4% of the participants in the CR group (77%). For the most difficult case, Case 1, 26% of participants in the SR group and 15% in the CR group reported guessing. On average, guessing was more often reported in the SR group (20% vs. 15%). Furthermore, participants in the CR group indicated lower levels of confidence ($M_{confCR} = 59.01$, $SD_{confCR} = 23.62$; $M_{confSR} = 67.04$, $SD_{confSR} = 23.16$) and reported the cases as being more difficult ($M_{diffCR} = 3.52$, $SD_{diffCR} = 0.96$; $M_{diffSR} = 3.21$, $SD_{diffSR} = 1.14$). Importantly, these descriptive statistics serve an illustrative purpose and should be interpreted accordingly. Hence, no significance testing was conducted.

A combined measure of task fluency

Generally, the within-person centered indicators for time on task, confidence and perception of difficulty were correlated to each other with Pearson correlations of $r_{(\text{time}*-1,\ \text{confidence})} = 0.48$, $r_{(\text{time}*-1,\ \text{difficulty})} = 0.62$, and $r_{(\text{difficulty}*-1,\ \text{confidence})} = 0.74$. All correlations were statistically significant with $p < 0.001$. In addition, we carried out a principal component analysis using Varimax rotation in the R package *psych* (Revelle, 2018). The results indicated that a common factor accounted for 74% of the variance in the observed variables. Furthermore, the correlation between the six case-level-averaged fluency measures and the six according case-level-averaged accuracies correlated with $r = 0.87$ ($t = 3.49$, $df = 4$, $p = 0.03$). Thus, the results were in line with our theoretical perspective and we summarized these measures into a single indicator of task fluency.

Antecedents of accuracy (generalized linear mixed effects model)

In order to address our research objectives we fitted successively more complex models using generalized linear mixed effects models. Across models, the random effects structure was identical. In the following, we highlight the main findings; details for the models are given in Table 2.

First, we only included the main effect for response format (i.e., the group effect) in the model. As expected, the result indicated that participants in the CR condition were less likely to give a correct diagnosis. The group effect alone explained 25% more of the variance on the between-person level as compared to the Null Model.

Second, we included both the main effect for the response format and CRT scores as fixed effects. Both were statistically significant predictors for giving a correct diagnosis and, combined, explained 43% additional variance on the between-person level ($OR = 0.50$, $p = 0.011$ and $OR = 1.58$, $p < 0.001$, respectively).

In a third step, we also included the fluency-related variables on the within-person level. This is, the person-specific variables task fluency and guessing that varied by cases. The results from this step indicate that, as expected, both a perception of low fluency and self-reported guessing were associated with decreased odds of giving a correct

Table 2 Results from the five different generalized mixed effects models

	Null Model		Model 1		Model 2		Model 3		Model 4	
	OR	*SE*	*OR*	*SE*	*OR*	*SE*	*OR*	*SE*	*OR*	*SE*
Fixed parts										
(Intercept)	1.79	0.93	2.59	1.41	0.90	0.55	0.97	0.57	0.46	0.65
Response format			0.47 *	0.15	0.50 *	0.15	0.47 *	0.31	1.84	0.67
Cognitive reflection					1.58 ***	0.22	1.65 ***	0.14	2.27 ***	0.21
Fluency							2.22 ***	0.21	2.21 ***	0.21
Guessing							0.51	0.38	0.50 *	0.38
Response format-cognitive reflection interaction									0.54 *	0.27
Random parts										
$\tau_{00,\text{ person}}$	0.522		0.392		0.169		0.160		0.065	
$\omega_{00,\text{ case}}$	1.428		1.432		1.455		1.085		1.083	
Variance explained per level (in relation to null model)										
Persons	–		25%		68%		66%		84%	
Cases	–		0%		−2%		29%		29%	

$*p<.05$, $**p<..01$, $***p<..001$

diagnosis ($OR=2.22$, $p<0.001$; $OR=0.51$, $p=0.079$, respectively). Both predictors combined explained 29% of the variance at the case level as compared to the null model.

Furthermore, we estimated a model introducing an interaction in order to investigate whether fluency was differently related to accuracy across the two conditions. The fluency-group interaction was not statistically significant (OR = 0.58; CI 0.27 – 1.25; $p=0.162$). The newly introduced interaction did not add explained variance on any of the levels (not reported in Table 2).

The final, and most complex model introduced a response-format-CRT interaction. In this model, higher scores on the CRT were associated with more than doubling the chances of a correct diagnosis ($OR=2.27$, $p<0.001$). This, however, was only the case for the SR group; the effect for the CRT-Response-Type-Interaction in the CR group was $OR=0.54$ ($p=0.026$), thus canceling out the CRT-main-effect within this group. This model accounted for 84% of the variance on the between-person level and 29% of the variance on the case-level. Finally, both fluency and guessing were case-specific, that is, introducing these predictors to the model did only account for little variance on the between-person level (1%).

The significant interaction and the amount of variance explained associated with CRT scores pointed to a substantial difference in the correlation between diagnostic accuracy and performance on the CRT across the two experimental groups. Indeed, the rank correlation (Spearman) between CRT score and number of correct cases within the SR condition was $r_{(\text{dx, crt})}=0.70$ as compared to $r_{(\text{dx, crt})}=-0.07$ within the CR group. We conducted a robustness check of this finding. Employing a bootstrap procedure, we found a 95% confidence interval for this correlation of 0.44 to 0.87 within the SR group. In the CR group, this interval was −0.05 to 0.34.

Predicting guessing

Finally, we fitted a GLMM in order to explain guessing on a given case. To do so, we basically exchanged two variables in the model; diagnostic accuracy became a predictor and guessing became the dependent variable. Furthermore, we included the person-centered fluency variable, and the CRT score to the predictors. The results indicated that guessing was less likely on cases where participants perceived high fluency ($OR = 0.29$; CI 0.17—0.49; $p < 0.001$). Neither the group-effect nor diagnostic accuracy, nor the CRT score showed a statistically significant relation ($OR = 0.58$, $p = 0.35$; $OR = 0.67$, $p = 0.34$; $OR = 1.35$, $p = 0.28$, respectively).

Discussion

In this article, we report an experimental study in which participants were randomly assigned to complete six clinical cases on shortness of breath in a clinical problem-solving scenario using either constructed response or selected response answering format. Based on our review of the literature, we approached two main research questions—the relation between accuracy and task fluency on the one hand, and the relation between cognitive reflection and accuracy on the other hand.

Similar to previous studies we found that participants were able to monitor their performance on a case (Eva & Regehr, 2007, 2011; Kämmer et al., 2020). Furthermore, we did not find an effect of the response condition on the relation between the perception of task fluency and accuracy. However, our most critical finding was that CRT scores were related to higher accuracy in the SR condition, but not in the CR condition. In addition, we found support for our expectation that guessing was related to the perception of low fluency. Interestingly, we did not find an advantage of guessing: when participants reported guessing, they were more likely to be incorrect in their diagnosis.

Overall, we interpret these results as supporting the stance that different response formats evoke different response behavior and, at the same time, pose different demands on test takers. On the group level, some markers of task fluency varied between the two response formats—for instance, and as expected, it took more time to answer cases in the CR format. At the same time, it appeared that discrepancies in accuracy between the two conditions were strongly related to scores on the Cognitive Reflection Test (i.e., we found a significant CRT-response format interaction). Given the fact that we used an experimental, randomized design and with the theoretical framework introduced in mind, the main contender for explaining these differences in scores between groups is the differential role of response inhibition in both conditions. Interestingly, an interaction between response-format and task fluency was not significant. While this doesn't allow for being interpreted as a null-effect, it still is interesting that the inhibition-response-format interaction was stronger than the interaction between fluency and response-format. We also want to mention that we did observe, as in many other studies in medical education, the phenomenon of 'case specificity'—correlations of accuracy between cases were ranging between $r = -0.5$ and $r = 0.28$. These were, indeed, in the range typically observed in clinical reasoning studies (Norman, 2008).

Studies focusing on the effect of response format in the assessment of clinical reasoning typically employ a theoretical framework rooted in dual-process theories (Monteiro

& Norman, 2013), making a clear distinction between fast and slow reasoning and their relation to success or failure on a case (Heemskerk et al., 2008). Drawing on recent research in cognitive psychology, we aimed at extending this approach using the concept of response inhibition and cognitive reflection. This framework suggests that both being able to detect a conflict within an intuitive response and to inhibit this intuitive response is critical to successful, that is, accurate, reasoning.

In this study, we could only provide indirect empirical support for this conclusion because the measure for inhibition—the Cognitive Reflection Test—was a distal indicator of this faculty. Indeed, to date, most research on response inhibition relies on such indicators—observing inhibition in vivo and on the level of particular cases would clearly require a more fine-grained approach. Nevertheless, the framework applied here raises several questions that have interesting implications for the broader field of assessment of clinical reasoning. For instance, it raises the issue of how switching to 'slow' reasoning really can be induced experimentally. It is an open question whether a simple instruction to engage in deliberate reasoning really affects a certain reflection on the initial response. Specific designs of study tasks and conditions are, indeed, an option rarely endorsed in medical education research–but are quite common in research in cognitive psychology.

The current study has several limitations. First of all, although we conducted a randomized trial, the findings might still be the result of the specific composition of the groups analyzed here. Building on a sample of N = 54 participants and six cases, our study was still comparable in size to similar experimental studies in the context of research on clinical reasoning. In general, statistically, more extreme effects are likely to be found in such smaller samples. Therefore, the correlation patterns found here should be interpreted with caution. We did, however, perform a non-parametric bootstrap to investigate in which respect the association between diagnostic accuracy and CRT scores was influenced by the specific composition of the groups investigated here. The bootstrap suggests that the effect found might be at the higher end of possible effect sizes. However, the 95% bootstrap confidence intervals did not overlap, which suggests that the finding of differential correlation patterns across groups is reasonably robust.

Furthermore, we opted for a between-person design so that there were sufficient replications within person. Consequently, there is no possibility of investigating possible interactions between persons and response format. Finally, results obtained here might not be readily transferable to high stakes contexts, such as in licensing exams. Such scenarios are usually characterized by higher psychological strain or stress which may have additional effects not observable in this study.

Our findings have practical implications, too. We argue that there is not one type of response format that is generally `more valid` than another for the specific purpose of assessing clinical reasoning. Fenderson and colleagues claim that multiple choice tests tend to focus on trivia (Fenderson et al., 1997), a position, in our experience, frequently raised by lecturers and professionals in medical education. While this might be true in some contexts, it is, obviously, not the response format alone that triggers the type of reasoning processes but rather the task as a whole. Indeed, while scores are assumed to be largely comparable, the cognitive process preceding the actual answer might not. Hence, we agree with Desjardins and colleagues' (Desjardins et al., 2014) conclusion that the exclusive use of only one type of response format might have unfavorable effects and could ultimately impair the validity of a test or assessment. In this respect, we propose that our findings support the stance that the design of assessments in medical education should aim for using heterogeneous response formats.

In conclusion, our study introduced a new theoretical account of how to characterize differences in task processing and investigated how different response formats relate to different task processing behaviors. We argue that the findings presented here support the stance that different response formats are related to different processing behavior. Consequently, when designing assessments of clinical reasoning, it is crucial to tap into different facets of this complex and important medical process.

Funding Open access funding provided by University of Oslo (incl Oslo University Hospital).

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*(4), 569–576. https://doi.org/10.1037/0096-3445.136.4.569

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

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *American Journal of Medicine, 121*(5 Suppl), S2-23. https://doi.org/10.1016/j.amjmed.2008.01.001

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55–68. https://doi.org/10.1037/0096-3445.127.1.55

Białek, M., & Sawicki, P. (2018). Cognitive reflection effects on time discounting. *Journal of Individual Differences, 39*(2), 99–106. https://doi.org/10.1027/1614-0001/a000254

Bleske-Rechek, A., Zeug, N., & Webb, R. M. (2007). Discrepant performance on multiple-choice and short answer assessments and the relation of performance to general scholastic aptitude. *Assessment & Evaluation in Higher Education, 32*(2), 89–105. https://doi.org/10.1080/02602930600800763

De Neys, W. (2013). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning, 20*(2), 169–187. https://doi.org/10.1080/13546783.2013.854725

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning, 20*(2), 169–187. https://doi.org/10.1080/13546783.2013.854725

De Neys, W., & Bonnefon, J. F. (2013). The "whys" and "whens" of individual differences in thinking biases. *Trends in Cognitive Sciences, 17*(4), 172–178. https://doi.org/10.1016/j.tics.2013.02.001

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition, 106*(3), 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

DeMars, C. E. (1998). Gender Differences in Mathematics and Science on a High School Proficiency Exam: The Role of Response Format. *Applied Measurement in Education, 11*(3), 279–299. https://doi.org/10.1207/s15324818ame1103_4

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55–77. https://doi.org/10.1207/s15324818ame1301_3

Desjardins, I., Touchie, C., Pugh, D., Wood, T. J., & Humphrey-Murto, S. (2014). The impact of cueing on written examinations of clinical decision making: A case study. *Medical Education, 48*(3), 255–261. https://doi.org/10.1111/medu.12296

Dunlosky, J., & Thiede, K. W. (2013). Metamemory. In D. Reisberg (Ed.), *The oxford handboog of cognitive psychology.* (pp. 283–298). Oxford University Press.

Elstein, A. S., Shulman, L. S., Sprafka, S. A., et al. (1978). *Medical problem solving: An analysis of clinical reasoning*. Harvard University Press.

Eva, K. W., & Regehr, G. (2007). Knowing when to look it up: A new conception of self-assessment ability. *Academic Medicine, 82*(10 Suppl), S81-84. https://doi.org/10.1097/ACM.0b013e31813e6755

Eva, K. W., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education: Theory and Practice, 16*(3), 311–329. https://doi.org/10.1007/s10459-010-9263-2

Evans, J. S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Fenderson, B. A., Damjanov, I., Robeson, M. R., Veloski, J. J., & Rubin, E. (1997). The virtues of extended matching and uncued tests as alternatives to multiple choice questions. *Human Pathology, 28*(5), 526–532. https://doi.org/10.1016/s0046-8177(97)90073-3

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

Heemskerk, L., Norman, G., Chou, S., Mintz, M., Mandin, H., & McLaughlin, K. (2008). The effect of question format and task difficulty on reasoning strategies and diagnostic performance in internal medicine residents. *Advances in Health Sciences Education: Theory and Practice, 13*(4), 453–462. https://doi.org/10.1007/s10459-006-9057-8

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*(5), 1191–1206. https://doi.org/10.1037/a0013025

Hickson, S., Reed, W. R., & Sander, N. (2012). Estimating the effect on grades of using multiple-choice versus constructive-response questions: Data from the classroom. *Educational Assessment, 17*(4), 200–213. https://doi.org/10.1080/10627197.2012.735915

Hift, R. J. (2014). Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Medical Education, 14*(1), 249. https://doi.org/10.1186/s12909-014-0249-2

Huwendiek, S., Reichert, F., Duncker, C., de Leng, B. A., van der Vleuten, C. P. M., Muijtjens, A. M. M., & Dolmans, D. (2017). Electronic assessment of clinical reasoning in clerkships: A mixed-methods comparison of long-menu key-feature problems with context-rich single best answer questions. *Medical Teacher, 39*(5), 476–485. https://doi.org/10.1080/0142159X.2017.1297525

Kämmer, J. E., Hautz, W. E., & März, M. (2020). Self-monitoring accuracy does not increase throughout undergraduate medical education. *Medical Education, 54*, 320–327. https://doi.org/10.1111/medu.14057

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119*(1), 80–113. https://doi.org/10.1037/a0025648

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22*(6), 787–794. https://doi.org/10.1177/0956797611407929

Kostopoulou, O., Rosen, A., Round, T., Wright, E., Douiri, A., & Delaney, B. (2015). Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients. *British Journal of General Practice, 65*(630), e49

Kostopoulou, O., Russo, J. E., Keenan, G., Delaney, B. C., & Douiri, A. (2012). Information distortion in physicians' diagnostic judgments. *Medical Decision Making, 32*(6), 831–839. https://doi.org/10.1177/0272989X12447241

Kostopoulou, O., Sirota, M., Round, T., Samaranayaka, S., & Delaney, B. C. (2016). The role of physicians' first impressions in the diagnosis of possible cancers without alarm symptoms. *Medical Decision Making, 37*(1), 9–16. https://doi.org/10.1177/0272989X16644563

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97–109. https://doi.org/10.1037/a0020762

Kunina-Habenicht, O., Hautz, W. E., Knigge, M., Spies, C., & Ahlers, O. (2015). Assessing clinical reasoning (ASCLIRE): Instrument development and validation. *Advances in Health Sciences Education: Theory and Practice, 20*(5), 1205–1224. https://doi.org/10.1007/s10459-015-9596-y

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology, 13*(3).

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*(3), 234–250

Mamede, S., Hautz, W. E., Berendonk, C., Hautz, S. C., Sauter, T. C., Rotgans, J., Zwaan, L., & Schmidt, H. G. (2020). Think twice: Effects on diagnostic accuracy of returning to the case to reflect upon the initial diagnosis. *Academic Medicine, 95*(8), 1223–1229. https://doi.org/10.1097/ACM.0000000000003153

Mamede, S., Splinter, T. A. W., van Gog, T., Rikers, R. M. J. P., & Schmidt, H. G. (2012). Exploring the role of salient distracting clinical features in the emergence of diagnostic errors and the mechanisms through which reflection counteracts mistakes. *BMJ Quality & Safety, 21*, 295–300

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207–218. https://doi.org/10.1207/s15326985ep3404_2

Monteiro, S. D., & Norman, G. (2013). Diagnostic reasoning: Where we've been, where we're going. *Teaching and Learning in Medicine, 25*(sup1), S26–S32. https://doi.org/10.1080/10401334.2013.842911

Monteiro, S. D., Sherbino, J. D., Ilgen, J. S., et al. (2015). Disrupting diagnostic reasoning: Do interruptions, instructions, and experience affect the diagnostic accuracy and response time of residents and emergency physicians? *Academic Medicine, 90*(4), 511–517. https://doi.org/10.1097/ACM.0000000000000614

Norman, G. R. (2008). The glass is a little full – of something: Revisiting the issue of content specificity of problem solving. *Medical Education, 42*, 549–551. https://doi.org/10.1111/j.1365-2923.2008.03096.x

Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education, 44*(1), 94–100. https://doi.org/10.1111/j.1365-2923.2009.03507.x

Norman, G. R., Swanson, D. B., & Case, S. M. (1996). Conceptual and methodological issues in studies comparing assessment formats. *Teaching and Learning in Medicine, 8*(4), 208–216. https://doi.org/10.1080/10401339609539799

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*(6), 237–241. https://doi.org/10.1016/j.tics.2008.02.014

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology, 67*(3), 215–227. https://doi.org/10.1037/a0032918

Patel, V. L., Groen, G. J., & Norman, G. R. (1993). Reasoning and instruction in medical curricula. *Cognition and Instruction, 10*(4), 335–378. https://doi.org/10.1207/s1532690xci1004_2

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). Everyday consequences of analytic thinking. *Current Directions in Psychological Science, 24*(6), 425–432. https://doi.org/10.1177/0963721415604610

R Core Team. (2018). R: A language and environment for statistical computing. From R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Revelle, W. (2018). psych: Procedures for Psychological, Psychometric, and Personality Research. Retrieved from https://CRAN.R-project.org/package=psych Version = 1.8.4.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163–184

Sam, A. H., Field, S. M., Collares, C. F., van der Vleuten, C. P. M., Wass, V. J., Melville, C., & Meeran, K. (2018). Very-short-answer questions: reliability, discrimination and acceptability. *Medical Education, 52*(4), 447–455. https://doi.org/10.1111/medu.13504

Schuwirth, L. W. T., Vleuten, C. P. M., & Donkers, H. H. L. M. (1996). A closer look at cueing effects in multiple-choice questions. *Medical Education, 30*(1), 44–49. https://doi.org/10.1111/j.1365-2923.1996.tb00716.x

Snijders, T. A. M., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. (2nd ed.). SAGE Publications.

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition, 128*(2), 237–251. https://doi.org/10.1016/j.cognition.2012.09.012

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory and Cognition*. https://doi.org/10.3758/s13421-011-0104-1

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

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the cognitive reflection test. *Cognition, 150*, 109–118. https://doi.org/10.1016/j.cognition.2016.01.015

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.