PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Can Clinical Case Discussions foster clinical reasoning skills in undergraduate medical education? A randomised controlled trial
AUTHORS	Weidenbusch, Marc; Lenzer, Benedikt; Sailer, Maximilian; Strobel, Christian; Kunisch, Raphael; Kiesewetter, Jan; Fischer, Martin; Zottmann, Jan

VERSION 1 – REVIEW

REVIEWER	Eugene Custers UMC Utrecht, Department of Educational Research and Development, Netherlands
REVIEW RETURNED	05-Nov-2018

GENERAL COMMENTS	Summary of work: The authors have performed a randomized controlled trial in which they compared three educational (instructional) approaches to learning clinical reasoning: participation in live clinical case discussions (CCD), watching videos of registered CCDs performed by others, or working with printed cases without discussing the case with others. Participants were medical students at a German university in their first through eigth clinical semester. The results showed a positive effect of CCD on a clinical skills reasoning tekst, both at immediate testing and, somewhat attenuated, at delayed testing. The authors interpret this as evidence that interactive formats (i.c., live-CCDs) are definitely more effective than paper cases, and probably also as videotaped discussions, in fostering students' clinical reasoning skills. General impression: I read the paper with interest and believe it could contribute to our knowledge about training clinical reasoning. However, I also found much of the necessary information to be lacking. Together with some questionable assumptions and practices, I find it hard to properly judge the soundness of the approach. Major issues: First, I do not wish to contend the conclusion that Live-CCDs resulted in higher scores on the knowledge test than Videotaped-CCDs, and Videotaped-CCDs in higher scores than Paper-CCDs. In this respect, the design of the study is sound. What I would want to challenge is whether this has anything to say about clinical reasoning. My default explanation would be that students who participate in the Live-CCDs just remember more from this educational experience than students in the other groups. As essential information about both the knowledge application/clinical reasoning test and about the content of the cases is missing, we cannot judge this. We learn that the test addresses conceptual,

	strategic, and conditional knowledge but it is unclear whether the authors constructed this test by themselves, or whether this instrument already has been used in earlier studies. No example questions are provided, so the reader cannot check what “strategic knowledge”, for example, would be. Surprisingly, we read in the article summary on page 3, second bullet from bottom: “The knowledge application test utilized in this study did not allow for a more indepth analysis of clinical reasoning skills (i.e., a distinction of conceptual, strategic, and conditional knowledge).” This issue is addressed nowhere in the remainder of the paper. About the cases, the only thing we learn is that they are “three independent internal medicine cases” (middle of page 7). Though the authors go to great length to demonstrate generalizability over participants (the checklist on page 24-25 attests to this), they do not discuss the issue of generalizability over cases (often called “transfer” in educational and psychological studies). This latter issue may be at least equally important in educational studies than randomization. For example, in a recent study performed by a Master-student at our institution (Keemink et al., 2018, International Journal of Medical Education, Vol 9, pp 35-41) she found that a positive effect of Case-Based Clinical Reasoning sessions – which may be somewhat similar to the approach studied in the present paper – did not extend to diseases of similar difficulty not addressed in the interactive sessions. I definitely do not wat to force the authors to refer to this study, but as they interpret their results in terms of clinical reasoning skills, the question of the generalizability of the effect over cases not included in the study should be discussed in a thoughtful manner. Finally, it is also not clear whether the discussant in the Live-CCD group had information about the case that was not available for the other participants (in particular, about the final diagnosis). If so, an interpretation of the results in terms of clinical reasoning skills would be further challenged, because then the reasoning process of the discussant could easily boil down to a reconstruction (including a possible hindsight bias) rather then “true” clinical problem solving. Ideally, to train “true” clinical reasoning, the leader of a clinical case discussion has the same information about the case, but more clinical experience in general, than the other participants. There are also a few other important issues with respect to the design of the study. For example, was the pretest identical to the knowledge application test (Table 1 suggests this, but information about the pretest is lacking). If this is the case, the results can be presented as the proportion of knowledge gained by the intervention (the difference between pretest and immediate posttest) which is retained at the delayed posttest. In addition, what was the time lag between T_0 and T_1 (Figure 2)? Was the pretest administered during the admission-part of the first of the three weekly CCDs, or in an individual session? Most importantly, did the students in the Paper-CCD spend the same amount of time on the cases than the students in the other groups? We learn there was no maximum time in the single-learning formats (second paragraph on page 9), but was there a minimum time? If not, the results may simply be the result of differences in time-on-task between the groups, in particular the Paper-CCDs. My gut feeling would tell me it is not easy to keep a student engaged for 75 minutes (the sum of the Discussion and Summary parts of the structure in Figure 1) on a single paper case. The possibility should at least be mentioned in the Limitations-subsection.

	Next, how did the randomization take place? Please note that if the randomization was properly carried out, it makes no sense to test for prior differences, because such differences would by definition be the result of chance (top of page 11). That is, it is not possible (it is even a fallacy) to check whether the randomization was "successfully" performed by investigating its ourcomes. In the first paragraph on page 7, we read that volunteers were "stratified by age, gender, year of study, prior CCD participation and performance in a knowledge application pre-test at T_0." I suppose stratification was performed before randomization, but in any case, the procedures should be accurately and completely described. By the way, my guess would be that "age" and "year of study" are strongly correlated; hence, it does not seem to make much sense to stratify on both variables. I also wonder whether it makes sense to stratify by gender, is there any apriori evidence of a gender difference on this type of task? The post-hoc testing of such differences (first paragraph on page 11) is questionable and the results, which nearly miss significance, suggest male students perform somewhat better on the test than female students, which probably is a matter of chance or may be explained by other features of the sample of participants (e.g., proportionally more male students from the more advanced clinical semesters). – In the first paragraph of page 11, the authors mention "drop-out", but this issue is not further pursued in the Results-section (I wonder where the "n=88" on the very first line of page 12 comes from). It would be very helpful, and I would definitely recommend, that the authors structure their Method-section in accordance with what is customary in educational/psychological studies, i.e., separate "Participants", "Design", "Materials", "Procedure", and "Analysis" subsections. A clear "Procedure" subsection (a detailed description of what happens to any individual participant from the time he/she signs up for participation until the final handshake) would be very helpful and maybe a "natural" way to answer most of the above questions. Minor issues: Whether a volunteer effect (referred to in the last bulleted point on page 3) actually occurred could be checked, e.g., by comparing examination results between participating and non-participating students. I find the range of students that were eligible for participation rather wide ("first to eighth clinical semester", page 7) rather wide. Was the effect of the experimental intervention predicted to be similar across the whole range? I suppose clinical reasoning skills of students in their eighth clinical semester are more advanced than those of students in their first semester, but this is not discussed. I prefer Cohen's d over eta-squared as a measure of effect size, as it provides a more immediate insight into the size of the effect. As the knowledge application test contained multiple choice questions, what would have been the minimum score (i.e., when a correction for guessing would be applied)? From the Results on page 11, I get the impression in the Paper-CCD group, some participants must have been unable to correctly answer any question at all. Even in the Live-CCD group, the effective maximum score appears to be around 20 (2 SDs above the mean) out of 29 correct answers. A short elucidation may be helpful.

	I wonder whether passive participation can be equated with "social loafing" (middle of page 5). I believe social loafing implies students' reaping benefits from collaborative activities without putting in efforts to justify these benefits. It occurs mostly when group performance is judged and the benefits (grades) are equally distributed over participants, regardless of any individual participant's efforts. If students are assessed individually, it is hard to see how passive participation can be considered social loafing: e.g., students who are passive in the group may have to study harder to pass the examinations, because they do not benefit as much from the learning experience than their more active peers, or they may be just more shy in these groups than their peers. Page 10, Statistical Analyses: I suppose where you mention "ANOVAs", you mean a ONEWAY analysis? If not, clarify. Last paragraph on page 12: Investigation of the relationship between subjective assessment and the results on the knowledge application test appears out of the blue here; it should be announced in the Introduction as part of the research question. Page 14, second paragraph: please note that the absence of significant differences between Live-CCD and Video-CCD groups is a result of post-hoc significance testing and should be interpreted with caution. Questions for clarification: Bottom of page 4, Merseth apparently requires cases to be "real". What does that mean? Is it prohibited to change a single detail in a case? And why would this be so? Top of page 5, what is the difference between a "case vignette" and an "elaborated and authentic case"? Can cases as used in this study (i.e., as topic of a live or videotaped discussion) be authentic at all, with no actual patient being involved? What is a "reasonable approach to the patient encounter" (page 8)? "… a clinician who could stop the discussion at any point when faulty reasoning was evident" (page 8): were any norms, or was there something like a blueprint (e.g., a predetermined acceptable line of reasoning) that was used to decide that "faulty reasoning" occured, or was this left to the discretion of the clinician?

REVIEWER	Odd Martin Vallersnes Associate professor Department of General Practice University of Oslo Norway
REVIEW RETURNED	03-Dec-2018

GENERAL COMMENTS	Dear authors Your manuscript describes a randomised controlled trial of a teaching intervention for medical students. The intervention is a clinical case discussion and you compare the impact on clinical reasoning skills with other case based learning interventions without the interactive discussion part. The intervention makes

	sense to try out based on previous research in the field. The study is well designed. Your findings have clear implications for how clinical reasoning could be taught. However, there are some issues that would benefit from some clarification. Methods Were the moderators recruited among the students volunteering for the study? If so, how were they chosen? If not, how were they recruited? Is the knowledge application test in your study a validated instrument? Has it been used previously or did you construct it yourselves? Was the knowledge application test identical at the three testing points? It would be good to show the questions used in the knowledge application test, as a table or as an appendix. Why was two weeks chosen as the time interval for the delayed knowledge application test? Results The difference between the Live-CCD and Video-CCD groups on the delayed knowledge application post-test was not statistically significant. Could this be due to underpowering of the study? Please comment in the Discussion section. Discussion You state that peer teaching courses might be easier to install and staff. I cannot quite see that this is so in your set-up, as an experienced clinician was present during the group discussions, and the moderators had to be trained. Please clarify. Minor issues Instead of listing just a few of the questions from the subjective learning outcomes questionnaire in the text, I would prefer to see all the questions in a table or in an appendix. Are the scores (M) for the knowledge application test in Results mean point scores (out of a maximum of 29)?

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: I do not wish to contend the conclusion that Live-CCDs resulted in higher scores on the knowledge test than Videotaped-CCDs, and Videotaped-CCDs in higher scores than Paper-CCDs. In this respect, the design of the study is sound. What I would want to challenge is whether this has anything to say about clinical reasoning. My default explanation would be that students who participate in the Live-CCDs just remember more from this educational experience than students in the other groups.	We appreciate the overall agreement of reviewer 1 with our conclusions. We assessed students' clinical reasoning skills via tests of knowledge application (not just a declarative knowledge test). We have now added more information in the Methods section to illustrate that our test instrument

	required more from the students than knowledge retention.
Reviewer 1: We learn that the test addresses conceptual, strategic, and conditional knowledge but it is unclear whether the authors constructed this test by themselves, or whether this instrument already has been used in earlier studies. No example questions are provided, so the reader cannot check what “strategic knowledge”, for example, would be.	The principle of knowledge application tests has been developed and validated by members of our workgroup (e.g. Kopp et al., 2009; Schmidmaier et al., 2013; Braun et al., 2017). This is stated more explicitly in the text now. Exemplary items from the test were added to the newly created figure 3.
Reviewer 1: Surprisingly, we read in the article summary on page 3, second bullet from bottom: “The knowledge application test utilized in this study did not allow for a more indepth analysis of clinical reasoning skills (i.e., a distinction of conceptual, strategic, and conditional knowledge).” This issue is addressed nowhere in the remainder of the paper.	Our knowledge application test includes items on conceptual (11 items), strategic (9 items), and conditional knowledge (9 items). Subscales for these knowledge types can generally be analyzed along with the overall test result as an indicator of clinical reasoning skills. While overall test reliability was satisfactory ($\alpha = .71$), larger item numbers would be necessary to reliably assess changes on the level of subscales. We have now added this point to the limitations.
Reviewer 1: About the cases, the only thing we learn is that they are “three independent internal medicine cases”. Though the authors go to great length to demonstrate generalizability over participants (the checklist on page 24-25 attests to this), they do not discuss the issue of generalizability over cases (often called “transfer” in educational and psychological studies). In a recent study performed by a Master-student at our institution (Keemink et al., 2018, International Journal of Medical Education, Vol 9, pp 35-41) she found that a positive effect of Case-Based Clinical Reasoning sessions – which may be somewhat similar to the approach studied in the present paper – did not extend to diseases of similar difficulty not addressed in the interactive sessions.	As indicated in the manuscript, all three cases (i.e. Kotton, Muse, & Nishino, 2012; Marks, & Zukerberg, 2004; Uyeki, Sharma, & Branda, 2009) were published in the New England Journal of Medicine. Chief complaints have been added to the description of the cases on page 8. We agree with reviewer 1 that clinical reasoning has been shown to be mostly domain-specific – a lack of specific conceptual knowledge in an area (e.g. the existence of certain diagnoses in a given medical field) obviously precludes successful diagnostic reasoning. While we did not investigate the transfer of clinical reasoning skills in this study, we feel that this is an important issue in the context of CCDs and have added this to the future research questions on page 17.

Reviewer 1: It is also not clear whether the discussant in the Live-CCD group had information about the case that was not available for the other participants (in particular, about the final diagnosis). If so, an interpretation of the results in terms of clinical reasoning skills would be further challenged, because then the reasoning process of the discussant could easily boil down to a reconstruction (including a possible hindsight bias) rather than "true" clinical problem solving. Ideally, to train "true" clinical reasoning, the leader of a clinical case discussion has the same information about the case, but more clinical experience in general, than the other participants.	Participants in all experimental groups were given identical information about all aspects of the cases (including the respective final diagnoses which were always disclosed at the end of a session). The CCD puts a strong focus on the peer teaching aspect; while an experienced clinician is present, he/she is not the leader, but just a supervisor of the case discussion. The presenter, the moderator, and the clinician are all facilitators of the discussion in the Live-CCD. We added more information on pages 9 and 10 to clarify this.
Reviewer 1: Was the pretest identical to the knowledge application test (Table 1 suggests this, but information about the pretest is lacking). If this is the case, the results can be presented as the proportion of knowledge gained by the intervention (the difference between pretest and immediate posttest) which is retained at the delayed posttest.	Meta-analyses on retest effects suggest that score increase is higher for identical forms than for parallel forms. In order to limit retest effects, we applied parallel forms of the knowledge application test (i.e., the topics covered by the individual items were the same, but the items were reformulated and their order permutated). We added this information on page 11.
Reviewer 1: What was the time lag between T_0 and T_1 (Figure 2)?	The time lag between T_0 and T_1 was four weeks. We have now added this on page 8.
Reviewer 1: Was the pretest administered during the admission-part of the first of the three weekly CCDs, or in an individual session?	The pretest (T_0) was administered in an introductory session one week before the weekly sessions started. We have added this on page 8.
Reviewer 1: Did the students in the Paper-CCD spend the same amount of time on the cases than the students in the other groups? We learn there was no maximum time in the single-learning formats (page 9), but was there a minimum time? If not, the results may simply be the result of differences in time-on-task between the groups, in particular the Paper-CCDs. My gut feeling would tell me it is not easy to keep a student engaged for 75 minutes (the sum of the Discussion and Summary parts of the structure in Figure 1) on a single paper case. The possibility should at least be mentioned in the Limitations-subsection.	There was neither a prespecified minimum nor a maximum time students were required to work on the cases in any group. This is now stated more explicitly on page 10. Indeed, we cannot entirely rule out time-on-task effects as suggested by reviewer 1 and have added this possibility to the limitations on page 17.
Reviewer 1: How did the randomization take place? Please note that if the randomization was properly carried out, it makes no sense to test for prior differences, because such differences would by definition be the result of chance (top of page 11). That is, it is not possible (it is even a fallacy) to check whether the randomization was	Randomisation was performed in a two-step procedure: First, we selected a sample of roughly 100 enrolled students. Next, we stratified this sample by creating triplets on the basis of the variables

“successfully” performed by investigating its outcomes. In the first paragraph on page 7, we read that volunteers were “stratified by age, gender, year of study, prior CCD participation and performance in a knowledge application pre-test at T_0”. I suppose stratification was performed before randomization, but in any case, the procedures should be accurately and completely described. By the way, my guess would be that “age” and “year of study” are strongly correlated; hence, it does not seem to make much sense to stratify on both variables.	age, gender, year of study, prior CCD participation, and performance in the knowledge application pre-test. This was done to limit the risk of random misdistribution of the selected sample. From each triplet we randomly assigned students to the experimental groups. We have added more information regarding the randomisation on page 7.
Reviewer 1: I wonder whether it makes sense to stratify by gender, is there any a priori evidence of a gender difference on this type of task? The post-hoc testing of such differences (first paragraph on page 11) is questionable and the results, which nearly miss significance, suggest male students perform somewhat better on the test than female students, which probably is a matter of chance or may be explained by other features of the sample of participants (e.g., proportionally more male students from the more advanced clinical semesters).	Female students outnumber male students in our medical school (ca. 70% of the students are female), so we anticipated a similar distribution within our sample and stratified for gender to avoid uneven distributions of male and female participants. Taking comments #12, #13 and #14 into account, we agree the “preliminary analyses” section is somewhat confusing and have therefore decided to remove it.
Reviewer 1: In the first paragraph of page 11, the authors mention “drop-out”, but this issue is not further pursued in the Results-section (I wonder where the “n=88” on the very first line of page 12 comes from).	In the Live-CCD, there was a dropout of 5 female participants. In the Video-CCD, 7 male participants dropped out. In the Paper-Cases group, 1 male and 3 female participants dropped out. As stated in the methods section, 90 (of 106) students completed the trial. Individual reasons for drop-out could not be obtained, but a 15% rate appears to be within normal range (e.g. Wood, White, & Thompson, 2004). We could not calculate subjective outcomes for 2 participants from the Paper-Cases group, as they had not fully answered the according questionnaire.
Reviewer 1: I recommend that the authors structure the Method-section in accordance with what is customary in educational/psychological studies, i.e., separate “Participants”, “Design”, “Materials”, “Procedure”, and “Analysis” subsections. A clear “Procedure” subsection (a detailed description of what happens to any individual participant from the time he/she signs up for participation until the final handshake) would be very helpful and maybe a “natural” way to answer most of the above questions.	We made an effort to adjust the methods section according to these suggestions.

Reviewer 1: Whether a volunteer effect (referred to in the last bulleted point on page 3) actually occurred could be checked, e.g., by comparing examination results between participating and non-participating students.	A comparison of subject-specific study examinations would be rather difficult (participants took these tests in different years, so data on the true comparison population would not be readily available). Participants had no influence on group allocation, so a volunteer effect would have affected all experimental groups evenly. Upon closer inspection, we feel this is not a central limitation of our study and have decided to remove that bullet point to avoid confusion.
Reviewer 1: I find the range of students that were eligible for participation ("first to eighth clinical semester", page 7) rather wide. Was the effect of the experimental intervention predicted to be similar across the whole range? I suppose clinical reasoning skills of students in their eighth clinical semester are more advanced than those of students in their first semester, but this is not discussed.	In fact, we hypothesized that participation in CCDs would increase clinical reasoning skills in all students, regardless of their clinical year. While we agree that clinical reasoning skills of more advanced students should be higher than those of beginners, the number of semesters studied is not a very reliable indicator for expertise. Having said that, there was no significant correlation of the semester number and the outcome variables in our data.
Reviewer 1: I prefer Cohen's d over eta-squared as a measure of effect size, as it provides a more immediate insight into the size of the effect.	We have now added Cohen's d effect sizes in addition to the partial eta-squared statistics.
Reviewer 1: As the knowledge application test contained multiple choice questions, what would have been the minimum score (i.e., when a correction for guessing would be applied)? From the Results on page 11, I get the impression in the Paper-CCD group, some participants must have been unable to correctly answer any question at all. Even in the Live-CCD group, the effective maximum score appears to be around 20 (2 SDs above the mean) out of 29 correct answers.	Individual test scores ranged from 1.5 to 21.5 across all three measurements (no participant scored 0 points). The difficulty of the knowledge application test was deliberately chosen to be high in order to avoid ceiling effects, as students from all clinical years were allowed to participate in the CCD as well as our study. We have added a sentence on the difficulty of the test on page 11.
Reviewer 1: I wonder whether passive participation can be equated with "social loafing" (middle of page 5). I believe social loafing implies students' reaping benefits from collaborative activities without putting in efforts to justify these benefits. It occurs mostly when group performance is judged and the benefits (grades) are equally distributed over participants, regardless of any individual participant's efforts. If students are assessed individually, it is hard to	We agree with the reviewer that the term "social loafing" might be misleading and rephrased the corresponding sentence on page 5 therefore.

see how passive participation can be considered social loafing.	
Reviewer 1: Page 10, Statistical Analyses: I suppose where you mention “ANOVAs”, you mean a ONEWAY analysis?	Correct, we modified the text on page 12 accordingly.
Reviewer 1: Last paragraph on page 12: Investigation of the relationship between subjective assessment and the results on the knowledge application test appears out of the blue here; it should be announced in the Introduction as part of the research question.	The investigation of subjective learning outcomes (in addition to measuring the students’ clinical reasoning skills) was added to the final paragraph of the introduction.
Reviewer 1: Page 14, second paragraph: please note that the absence of significant differences between Live-CCD and Video-CCD groups is a result of post-hoc significance testing and should be interpreted with caution.	See also comment #32 by reviewer 2 who suggested the Live-CCD and Video-CCD did not differ in the delayed post-test due to underpowering. We have inserted a sentence to clarify that this particular finding has to be treated with caution.
Reviewer 1: Bottom of page 4, Merseth apparently requires cases to be “real”. What does that mean? Is it prohibited to change a single detail in a case? And why would this be so? Top of page 5, what is the difference between a “case vignette” and an “elaborated and authentic case”? Can cases as used in this study (i.e., as topic of a live or videotaped discussion) be authentic at all, with no actual patient being involved?	Indeed, realism and authenticity are varying features of cases. The definition from Merseth’s review suggests that a case should be “based on a real-life situation or event”. We have now rewritten the according passage in the introduction.
Reviewer 1: What is a “reasonable approach to the patient encounter” (page 8)? “… a clinician who could stop the discussion at any point when faulty reasoning was evident” (page 8): was there something like a blueprint (e.g., a predetermined acceptable line of reasoning) that was used to decide that “faulty reasoning” occured, or was this left to the discretion of the clinician?	The “reasonable approach to the patient encounter” refers to the implementation of clinical rules such as “start with non-invasive tests, before you do invasive ones” or the notion that it will take 2-3 days for blood cultures to be reported positive, while results of a CT scan can be obtained within hours of admission. The involvement of the clinician into the discussion was left at his or her discretion. We updated the description on pages 9 and 10.
Reviewer 2: Were the moderators recruited among the students volunteering for the study? If so, how were they chosen? If not, how were they recruited? (Methods)	CCD moderators were/are recruited among previous CCD participants. Within their first year of moderating CCDs, these students participate in a training weekend, where they learn about some background in higher education and group facilitation. We updated the part on moderators on pages 9 and 10.

Reviewer 2: Is the knowledge application test in your study a validated instrument? Has it been used previously or did you construct it yourselves? (Methods)	The principle of knowledge application tests has previously been published and validated by members of our workgroup (see also our response to comment #4 by reviewer 1).
Reviewer 2: Was the knowledge application test identical at the three testing points? (Methods)	No, it was not identical. We used parallel forms of the test in order to limit retest effects (see also our response to comment #8 by reviewer 1).
Reviewer 2: It would be good to show the questions used in the knowledge application test, as a table or as an appendix. (Methods)	Exemplary items from the knowledge application test were added to the newly created figure 3.
Reviewer 2: Why was two weeks chosen as the time interval for the delayed knowledge application test? (Methods)	While many clinical reasoning studies conduct the delayed post-test after one week, we deliberately chose a slightly larger interval to investigate the sustainability of effects. We have now included a sentence on this on page 8.
Reviewer 2: The difference between the Live-CCD and Video-CCD groups on the delayed knowledge application post-test was not statistically significant. Could this be due to underpowering of the study? Please comment in the Discussion section. (Results)	The absence of a significant difference between the Live-CCD and Video-CCD groups in the delayed post-test could be due to underpowering, yes. We designed our trial to have a power of 80% for detecting a medium effect size. We are now acknowledging this possibility in the discussion on page 15.
Reviewer 2: You state that peer teaching courses might be easier to install and staff. I cannot quite see that this is so in your set-up, as an experienced clinician was present during the group discussions, and the moderators had to be trained. Please clarify. (Discussion)	We agree that because of the presence of an experienced clinician the logistic/economic benefit of the CCD format might be smaller compared to other peer teaching courses, but we still see some advantages. We have modified our statement on page 16 accordingly.
Reviewer 2: Instead of listing just a few of the questions from the subjective learning outcomes questionnaire in the text, I would prefer to see all the questions in a table or in an appendix.	All items from the subjective learning outcomes scale were added to the supplementary file.
Reviewer 2: Are the scores (M) for the knowledge application test in Results mean point scores (out of a maximum of 29)?	Yes (see also our response to comment #20 by reviewer 1).

VERSION 2 – REVIEW

REVIEWER	Eugene Custers UMC Utrecht, The Netherlands
REVIEW RETURNED	11-Apr-2019

GENERAL COMMENTS	I believe the study is well designed and neatly performed. However, I still differ with the authors on the interpretation of the results. I believe the study has little to say about clinical reasoning per se, it says something about knowledge acquisition. Table 1 is perfectly clear in this respect. We see a knowledge increase from pre-test to posttest (the size of which differs significantly between the three groups) followed by a relatively small decline over the course of two weeks, from posttest to delayed test (maybe this decrease, which would - rough estimation - amount to 8% in the Live-CCD group. 0% in the Video-CCD group, and 22% in the Paper-cases group) is not even significant, or only significant for the Paper-Cases group. Testing for between-group differences at the posttest or delayed test doesn't make much sense to me, it is the decrease over the retention interval that counts (this can tested with a more powerful within-subjects test). In short, to me this is basically a knowledge retention study, rather than a study on clinical reasoning. Some minor comments: page 4, line 23 (in margin): please check whether "amongst" is correct; page 5, lines 25-30, "any intervention that...": this seems to me a tautology (i.e., whether some cognitive activity is effective is assessed by improved learning outcomes, there is no independent way to do this); page 6, line 21: "cases are presented only until the hospital admission of the patient" is ambiguous; page 12, last paragraph: how should I interpret the Eta-Square (or Cohen's d) in a comparison between three groups?; similarly in the last paragraph on page 13; page 16, first sentence: What is the subject of "and were linked to their performance..." - I do not understand this sentence; page 16, line 32: "special preparation is not necessary…" is this really true? Can a clinician provide good quality teaching without doing any special preparation?

REVIEWER	Odd Martin Vallersnes Department of General Practice University of Oslo Norway
REVIEW RETURNED	20-Mar-2019

GENERAL COMMENTS	I am happy with the changes made in the revised manuscript.

VERSION 2 – AUTHOR RESPONSE

#	Comment	Actions taken
1	Reviewer 1: I believe the study has little to say about clinical reasoning per se, it says something about knowledge acquisition. Table 1 is perfectly clear in this respect. We see a knowledge increase from pre-test to posttest (the size of which differs significantly between the three groups) followed by a relatively small decline over the course of two weeks, from posttest to delayed test (maybe this decrease, which would - rough estimation - amount to 8% in the Live-CCD group. 0% in the Video-CCD group, and 22% in the Paper-cases group) is not even significant, or only significant for the Paper-Cases group. Testing for between-group differences at the posttest or delayed test doesn't make much sense to me, it is the decrease over the retention interval that counts (this can tested with a more powerful within-subjects test). In short, to me this is basically a knowledge retention study, rather than a study on clinical reasoning.	Apparently, reviewer 1 has a slightly different view on clinical reasoning than we do. As we see it, clinical reasoning skills can be measured effectively via tests of knowledge application (that incorporate key feature problems on “how information?” and problem-solving tasks on “why information?”). We describe our operationalization of clinical reasoning skills as instances of knowledge application in detail in the methods section. Consequently, we think that our post-test results say more about (the sustainability of) knowledge application skills than knowledge retention only. Undoubtedly, our test also assesses knowledge retention which does not contradict our arguments to use such a test to assess important aspects of clinical reasoning skills. We do, however, agree with the reviewer that there are other important facets of the clinical reasoning process that we did not assess. We have therefore added this aspect as a limitation of our study.
2	Reviewer 1: page 4, line 23 (in margin): please check whether "amongst" is correct.	We have changed the wording to "among".
3	Reviewer 1: page 5, lines 25-30, "any intervention that...": this seems to me a tautology (i.e., whether some cognitive activity is effective is assessed by improved learning outcomes, there is no independent way to do this).	We agree and have removed the sentence "Based on the ICAP model, any intervention that would lead to more effective cognitive (i.e. constructive or interactive) learner activities should improve the learning outcomes of that format".
4	Reviewer 1: page 6, line 21: "cases are presented only until the hospital admission of the patient" is ambiguous.	We made an effort to further clarify the timeline of the CCD process on page 6 (in which the admission to the hospital usually is the starting point for the group discussion phase).

5	Reviewer 1: page 12, last paragraph: how should I interpret the Eta-Square (or Cohen's d) in a comparison between three groups?; similarly in the last paragraph on page 13.	We feel that the addition of Cohen's d may be confusing (it is usually reported for a comparison of two groups), so we have now removed it from pages 12 and 13. Partial eta squared is used specifically in ANOVA models and is the default effect size measure in SPSS for ANOVA procedures. According to Richardson (2011), a value of .01 can be interpreted as a small effect, .06 as a medium effect, and .14 as a large effect.
6	Reviewer 1: page 16, first sentence: What is the subject of "and were linked to their performance..." - I do not understand this sentence.	We have changed the wording to "Subjective learning outcomes suggest that students prefer the live discussion over the other formats. The subjective assessment correlated with the students' performance in both knowledge application post-tests."
7	Reviewer 1: page 16, line 32: "special preparation is not necessary…" is this really true? Can a clinician provide good quality teaching without doing any special preparation?	We have added a sentence on page 16 to further clarify the role of the facilitating experienced clinician in the CCD approach.