PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Assessing the effect of empathy-enhancing interventions in health education and training: A systematic review of randomised controlled trials
AUTHORS	Winter, Rachel; Issa, Eyad; Roberts, Nia; Norman, Robert; Howick, Jeremy

VERSION 1 – REVIEW

REVIEWER	Caroline Wellbery Georgetown University Medical Center, USA
REVIEW RETURNED	09-Feb-2020

GENERAL COMMENTS	Context: the findings of this study are consistent with other studies, basically affirming that in spite of the heterogeneity of studies, there is evidence that interventions can cultivate empathy. The rest is nuance, and that brings me to the problem of such studies: what can another systematic review to our understanding of the various dimensions of empathy and its impacts? To this end, the author will have to update the intro with the recently posted review & meta-analysis of empathy interventions in medical students--Fragkos and Paul "The effectiveness of teaching Clinical Empathy to Medical Students: A systematic Review and Meta-analysis of RCT's" Academic Medicine October 2019 Overall study design and content. Full disclosure: I am far from having any expertise in the evaluation of systematic review design. However, the authors seem to have done their homework. For me, the most valuable portion of this manuscript was actually e-Table 3, describing the characteristics of the studies, which contained not only useful insights into study design but also provided details on the interventions used in the studies. However, the purpose of a review such as this one is not to teach study design (most readers would not focus on that) but rather what the practical take-homes are for application of the authors' findings and remaining knowledge gaps. So my recommendation to these authors is to streamline the description of methods (there is a lot of redundancy in the text and the e-content) and focus more, especially in the results and conclusions, on what kind of useful information the study now adds, in practical fashion, to educators. That, in my view, is a great strength of the Patel study the authors cite, which advised that based on their findings, educators should emphasize rehearsal and practice of specific behaviors (even if not everyone would be comfortable with that approach).

Specifics

P, 5 Ll 6-11 No standard empathy training exists: Does this paper solve this problem? If not, connect with a rationale that justifies undertaking a systematic review
Three systematic reviews: note new reference mentioned above

ll. 20-23 “interventions can increase empathy” Given that the authors talk about empathy decline, do they mean increase empathy of stave off empathy decline? Without specifics, it’s unclear why empathy decline is mentioned previously. Do any of the studies reviewed actually target empathy decline?

p. 7 ll. 3-6 It’s unclear to me how identifying the tools relates to the effectiveness of the interventions. Do the authors mean that you need reliable tools to have reliable measurement?

55-57 “Scores…were included” In what sense included? Do the the authors mean that in addition to other evidence of clinical empathy, empathy scores were included as evidence? (e.g. with the implications that score outcomes were considered valid evidence). Needs clarification

p. 10 ll. 8-1: If a study used different scales…primary outcome. For the non-researcher, an example would help, because the reader may not automatically understand the connection between scales and outcome. See comment below re primary outcome measures

p. 12 ll. 28-31 serious gaming interventions—why plural, was there more than 1?

p. 18 ll 13-18. Blinding…I think the reader needs to have the reason why blinding wasn’t possible made more explicit. For example, when people are filling out questionnaires, is it because they are not blind to the topic (empathy) at hand? Is this an issue of social desirability bias?

p. 20 ll. 49-56 The mention of the sub-group analysis and then the meta-analysis looking at improved empathy over time is confusing. Is this meta-analysis that is mentioned of the 11 studies? Maybe the mention of the sub-group analysis can be skipped, as the reader wonders what that analysis is, but doesn’t want to have to look into an e-supplement.

p. 21 ll 18-23. This is curious as the Fragkos and Paul systematic review found the greatest effect for this group. Might need explanation for those looking at both reviews.

Ll22-23 Arts and humanities: one other study had found low effects for arts and humanities. Is this related in any way to the number of studies? After all, the communication interventions were plentiful and also showed the greatest effects

p 23. ll 9-17 and 32-34: Ok for conclusion of a moderate effect, but saying the effects are well ‘sustained over time’ seems misleading—who decides that 12 weeks is enough? It seems that there was only a small effect after 12 weeks, which I would interpret as meaning in the best case scenario the interventions’

	effects attenuate over time. As a reader, I'd want more justification for the emphasis on sustained effect. It seems as though interventions of longer duration were more effective: why wouldn't this be highlighted, as it seems more convincing that sustained effect? Also, though not a goal of this study, it would seem interesting to look at the interaction between duration and sustained effect. Ll 29-31 Of note, the Fragkos review requires that the author update statements about what this review adds to the literature Ll 46-48 Wouldn't the quality of the review—or rather its conclusions-- be affected by the low quality of evidence as stated in the abstract? I'm not sure if the authors are making a distinction between quality of reporting and quality of evidence. Both strengths and limitations are cursory and need further fleshing out. Please continue to be explicit about primary outcome measures=assessment tools. If that is what is meant, it is still confusing when the authors mention the first reported measure of empathy. A little clarification could go a long way. p. 24 Implications for research and practice This is the crux of the review and yet this paragraph says nothing of substance. Here is the opportunity for the authors to say what this review offers that can lead to changes in practice. Telling investigators they need to develop robust interventions may be true, but is so general as not to be helpful Conclusion 46-53 How would investigators use this review? The authors need to use this opportunity to spell out the raison d'etre of their study Table 1 The Gholamzadeh row: there is no entry in the effect of intervention column A number of typos

REVIEWER	Marco Antonio de Carvalho Filho University Medical Center Groningen, The Netherlands
REVIEW RETURNED	09-Apr-2020

GENERAL COMMENTS	Dear editor, Thank you for the opportunity to review the article bmjopen-2019-036471 "Assessing the effect of empathy-enhancing interventions in health education and training: A systematic review of randomised controlled trials". The subject is very relevant for the medical education field and offers a starting point for researchers interested in exploring the field. The overall quality of the article is good and some improvements can be easily achieved by the authors. I hope the commentaries below are helpful. 1 – In the introduction the authors state the importance of empathy for clinical care, they succinctly describe the core elements of empathy as a concept and share the previous attempts of

	reviewing the efforts to teach empathy. However, the discussion about how to measure empathy, and what empathy means in the clinical x non-clinical context is missing. Considering that the authors review focus on measuring empathy after specific interventions, the reader would benefit from being introduced to how empathy is or can be measured in the medical education field. 2 – Although the effort of authors in reviewing the literature and synthetizing the data is valuable, the articles are very heterogeneous, and we cannot be sure that these articles are measuring the same construct. So, I suggest the authors a separate analysis for the articles that are using the same kind of measurements, for instance CARE or JSE, which seems to be the most used, with the first focusing in self-assessment and the later on patient assessment. 3 – The discussion starts with a statement that the interventions were efficient in improving empathy, but the authors are assuming that self-reported empathy is an accurate measurement of empathy, and this is debatable. I think the authors should better elaborate on that, and without discussing what each one of the scales are measuring, it is impossible to be faithful to the findings and not misguide the reader.

REVIEWER	Behal CHU Lille France
REVIEW RETURNED	28-May-2020

GENERAL COMMENTS	Comment for the editor. I have carefully reviewed the paper entitled « Assessing the effect of empathy-enhancing interventions in health education and training: A systematic review of randomised controlled trials. » submitted to Open BMJ by Rachel Winter et al. The present manuscript assess effect of empathy-enhancing interventions among student in the health field by performing a systematic review and meta-analysis. The manuscript was well written, the systematic review and meta-analysis was well done, following the PRISMA guideline with details on studies selection. As underlined by authors, a high heterogeneity was anticipated and observed likely due to the low quality of available RCT. The statistical section is well written, with a typo mistake in p-value to declare heterogeneity "(statistically significant for p0.01)"; I believe that authors wanted to write "heterogeneity was declared if p-value for heterogeneity<0.10 or I2>50%. There is some inconsistency between results section and flowchart (Figure 1) for the number of articles retrieved for full-text review (73 in the flowchart vs 72 in the text) and for the number of studies excluded (47 in the flowchart, 46 in the text). Please clarify. Similarly, there is some inconsistency between results section and others figures/tables: For example, in the result section you state that the duration of intervention ranges from 20 minutes to 18 hours, however in the Table 1, I found a duration of 42 hours for Yang's study. In the table 2, the heterogeneity for the study with least risk of bias is 66% whereas in the Figure 3 it is of 63%. There is also inconsistencies between values of standardised mean difference and their 95% confidence interval in the Table 2 and the supplemental Figures for effect of duration of intervention and participant population.

	I suggest also to express results with the same number of decimal. For example, in the first line of the Study design section, the number of participant could be expressed as entire value: median of 90 and IQR (49-154) In the eTable 4, there is a typo mistake in the upper limit of the 95% CI of the risk with empathy training SMD, the value should be 0.67 instead of 0.37?

VERSION 1 – AUTHOR RESPONSE

Comments from the reviewer	**Authors Reply**
Reviewer #1	
General comments	
Context: the findings of this study are consistent with other studies, basically affirming that in spite of the heterogeneity of studies, there is evidence that interventions can cultivate empathy. The rest is nuance, and that brings me to the problem of such studies: what can another systematic review to our understanding of the various dimensions of empathy and its impacts? To this end, the author will have to update the intro with the recently posted review & meta-analysis of empathy interventions in medical students-- Fragkos and Paul "The effectiveness of teaching Clinical Empathy to Medical Students: A systematic Review and Meta-analysis of RCT's" Academic Medicine October 2019	We thank the reviewer for drawing our attention to review. We have now added this reference to our manuscript, detailing in the introduction what this review adds and how our review differs: *"*Frakgos and Paul[22]conclude that empathy interventions significantly increase empathy, but limit their study population to medical students only. In addition they do not explore whether any improvement in empathy is sustained over time*"* PAGE 5
Overall study design and content. Full disclosure: I am far from having any expertise in the evaluation of systematic review design. However, the authors seem to have done their homework. For me, the most valuable portion of this manuscript was actually e-Table 3, describing the characteristics of the studies, which contained not only useful insights into study design but also provided details on the interventions used in the studies.	We thank the reviewer for this comment.

However, the purpose of a review such as this one is not to teach study design (most readers would not focus on that) but rather what the practical take-homes are for application of the authors' findings and remaining knowledge gaps. So my recommendation to these authors is to streamline the description of methods (there is a lot of redundancy in the text and the e-content) and focus more, especially in the results and conclusions, on what kind of useful information the study now adds, in practical fashion, to educators. That, in my view, is a great strength of the Patel study the authors cite, which advised that based on their findings, educators should emphasize rehearsal and practice of specific behaviors (even if not everyone would be comfortable with that approach).	This is an excellent suggestion that we have used to improve the manuscript. We have read the comments and suggestions provided by reviewer #1 carefully and have used these to guide amendments to our review. We have extended our 'discussion' section, specifically focusing on the expanding the 'summary of evidence', 'strengths and limitations' and 'implications for research and practice'. As suggested by reviewer #1 we have streamlined the description of methods and moved some further material to the eMethods section in the supplement. See responses to comments below for more details.
Specific comments	
P, 5 Ll 6-11 No standard empathy training exists: Does this paper solve this problem? If not, connect with a rationale that justifies undertaking a systematic review Three systematic reviews: note new reference mentioned above	This is a fair comment and in response we have added further clarification at the end of the introduction as to why we feel a further systematic review is justified. "No standard empathy-curriculum for healthcare training currently exists and empathy-based training does not appear routinely in healthcare education.[14] Understanding what type of empathy training is most effective in healthcare at both cultivating and sustaining empathy would be a useful start in preparing one." Page 4
ll. 20-23 "interventions can increase empathy" Given that the authors talk about empathy decline, do they mean increase empathy of stave off empathy decline? Without specifics, it's unclear why empathy decline is mentioned	Empathy decline is discussed in the introduction to provide context for the increased interest in empathy training for medical and healthcare students. We accept the reviewer's comments here and have amended the introduction to read:

previously. Do any of the studies reviewed actually target empathy decline?	"Although contested by some,[12,13] there is evidence that empathy in medical and health science students declines during undergraduate education.[14-16] Researchers agree that empathetic skills can be taught [17-20] and cultivating empathy to protect against a possible decline would seem sensible." Page 4
p. 7 ll. 3-6 It's unclear to me how identifying the tools relates to the effectiveness of the interventions. Do the authors mean that you need reliable tools to have reliable measurement?	We agree with the reviewer that this is unclear. We have amended this objective to read: "to identify the tools used to measure empathy levels in participants to consider differences in self-reported and observer-reported measures" Page 6
55-57 "Scores…were included" In what sense included? Do the the authors mean that in addition to other evidence of clinical empathy, empathy scores were included as evidence? (e.g. with the implications that score outcomes were considered valid evidence). Needs clarification	We agree with the reviewer that this is unclear and have amended this sentence to read: "Trials measuring empathy via self- and/or observer-reported measures were included." Page 7
p. 10 ll. 8-1: If a study used different scales…primary outcome. For the non-researcher, an example would help, because the reader may not automatically understand the connection between scales and outcome. See comment below re primary outcome measures	We have amended this sentence to provide further clarity. "If a study provided measures of empathy using different tools, the primary tool to measure empathy was used. If it was unclear which was the primary measure, we used the first reported measure of empathy." Page 8-9
p. 12 ll. 28-31 serious gaming interventions—why plural, was there more than 1?	We have amended this to reading 'serious gaming' rather than 'serious gaming interventions'. Page 11

p. 18 ll 13-18. Blinding…I think the reader needs to have the reason why blinding wasn't possible made more explicit. For example, when people are filling out questionnaires, is it because they are not blind to the topic (empathy) at hand? Is this an issue of social desirability bias?	We have added further clarification to this section in the 'risk of bias within studies': "Blinding was not possible in the majority of studies due to the nature of the interventions (often described to participants as empathy-promoting) and the method of outcome assessment (for example a self-report questionnaire which makes explicit what is being measured, such as the JSE)." Page 15 Further information about risk of bias within studies and blinding is reported in the eResults.
p. 20 ll. 49-56 The mention of the sub-group analysis and then the meta-analysis looking at improved empathy over time is confusing. Is this meta-analysis that is mentioned of the 11 studies? Maybe the mention of the sub-group analysis can be skipped, as the reader wonders what that analysis is, but doesn't want to have to look into an e-supplement.	We have amended this sentence to reduce confusion and the 'Sustainability of improved empathy analysis' section now reads: "Eleven studies provided follow-up data assessing sustainability of changes to empathy, in addition to post-intervention measurement. [23,28,29,31,33,35,37,39,41,49,52] Eight were eligible for inclusion in a sub-group analysis [23,29,35,37,39,41,49,52] (see eResults for further details) which found a moderate effect size for sustainability up to 12 weeks and a smaller, but still significant effect size for sustainability of impact of training at 12 weeks or later (figure 4 and table 2). Page 18
p. 21 ll 18-23. This is curious as the Fragkos and Paul systematic review found the greatest effect for this group. Might need explanation for those looking at both reviews.	This is interesting and the differences may be explained by: • The study populations for this review and Fragkos and Paul's review are different (Fragkos and Paul look only at a medical student population) and so the studies involved in the meta-analysis are different. • Fragkos and Paul report the largest effect for 'mixed' educational programmes but report very high heterogeneity. The trials included for

	the 'type of intervention' sub-group analysis in our review report a smaller (but significant) effect but heterogeneity between studies is much lower.
Ll22-23 Arts and humanities: one other study had found low effects for arts and humanities. Is this related in any way to the number of studies? After all, the communication interventions were plentiful and also showed the greatest effects	This is a very reasonable point and we have added a sentence to the results section to acknowledge this: "The smallest effect reported was for interventions that were described as 'mixed educational programmes' and ones based in the arts and humanities (table 2). It is worth noting however that only two studies used arts and humanities interventions (compared to seven in the communications skills group) and this may well impact on the effect size." Moreover, empathy studies do not always involve arts and humanities components. Page: 19
p 23. ll 9-17 and 32-34: Ok for conclusion of a moderate effect, but saying the effects are well 'sustained over time' seems misleading—who decides that 12 weeks is enough? It seems that there was only a small effect after 12 weeks, which I would interpret as meaning in the best case scenario the interventions' effects attenuate over time. As a reader, I'd want more justification for the emphasis on sustained effect.	We agree with the reviewer that 'sustained over time' could be viewed as misleading and have taken these recommendations into account. We have re-written the 'summary of evidence' to provide further clarity: "Training healthcare practitioners and trainees improved their empathy by a modest amount. The effect of training seemed to diminish, but lasts to beyond 12 weeks." Page 20
It seems as though interventions of longer duration were more effective: why wouldn't this be highlighted, as it seems more convincing that sustained effect? Also, though not a goal of this study, it	We thank the reviewer for this comment. Whilst not within the scope of this study, it would be interesting to look at the interaction between duration and sustained

would seem interesting to look at the interaction between duration and sustained effect.	effect. We have highlighted that longer duration interventions are more effective in the results section. Page 19
Ll 29-31 Of note, the Fragkos review requires that the author update statements about what this review adds to the literature	We have used the constructive comments from reviewer #1 to re-write the 'comparison with other evidence' section and have referenced the Fragkos and Paul review: Our review supports the evidence of previous similar reviews, finding benefits of empathy training[17,20,21,22] and that practitioner empathy training makes a difference to patients.[59] Our study adds to this evidence by providing an estimate of empathy training from higher quality (randomised) trials, and by showing that the effect lasts well beyond the intervention." Page 20
Ll 46-48 Wouldn't the quality of the review—or rather its conclusions-- be affected by the low quality of evidence as stated in the abstract? I'm not sure if the authors are making a distinction between quality of reporting and quality of evidence.	We thank the reviewer for drawing our attention to this and have re-written to sentence: "Also, the strength of findings in this review may be limited by the reporting quality of some of the included studies." Page 21
Both strengths and limitations are cursory and need further fleshing out.	We have followed this suggestion (and those of reviewer #2) and expanded the strengths and limitations section: "This review, to the best of our knowledge, is the first systematic review and meta-analysis limited to RCTs of clinical empathy training for all healthcare students and professionals. This is an up-to-date review that excludes non-randomised studies, follows a pre-published protocol and assesses both the immediate and longer term effects of empathy training. Our broad study population, with both healthcare students and

	professionals means findings are generalisable to all areas of healthcare education and training. We chose to include only the results of the primary measure of empathy reported by each study. Where it was unclear which was the primary measure, we used the measure that was reported first. We recognise that this might have been biased, as authors may have chosen to report the most positive outcomes first. However, we found that this was not necessarily the case. For example, the first measure of empathy reported by Buffel du Vaure et al [30] (who did not specific which measure was primary) had a smaller effect than the second. We recognise the heterogeneity of the studies in our review and anticipated this. This means that further research is required to identify the most effective empathy training methodology. Also, the strength of findings in this review may be limited by the reporting quality of some of the included studies. A sensitivity analysis of studies of highest quality found a slightly smaller but still significant effect size. Another limitation in reviewing the evidence in this field is the multiple tools used by investigators to measure clinical empathy. With the lack of a definitive definition of clinical empathy and a range of tools measuring different aspects of empathy, the impact of an intervention may vary depending on the measurement tool used. This is demonstrated by Reiss et al [44] who found a statistically significant improvement in empathy when measured using the CARE scale but no significant changes using the JSE. In contrast Buffel du Vaure [30] reported the opposite. Perhaps because of the larger sample size or other factors, our review found a benefit of training independently of how it was measured. A further limitation with this review is that we only identified four studies that followed participants up for at least three months. The trials identified however found a positive effect. Lastly, we did not measure the qualitative experiences of participants in this review.” Page 20-22

Please continue to be explicit about primary outcome measures=assessment tools. If that is what is meant, it is still confusing when the authors mention the first reported measure of empathy. A little clarification could go a long way.	We have noted this comment and ensured that we have referred to assessment tools when referring to primary outcome measures.
p. 24 Implications for research and practice This is the crux of the review and yet this paragraph says nothing of substance. Here is the opportunity for the authors to say what this review offers that can lead to changes in practice. Telling investigators they need to develop robust interventions may be true, but is so general as not to be helpful	We agree with the reviewer comments for the 'implications for research practice' and thank them for drawing our attention to this. We have re-written this section to read: "Interventions for cultivating student and trainee empathy should be further developed and implemented. Optimizing implementation will require additional qualitative research on the experiences of empathy teachers and learners. Also, the longer term effects (>12 weeks) of empathy training has not been studied adequately and future research should address this. With competition for time and space in both undergraduate and postgraduate healthcare curriculums, future research in this area needs to be robust. Designers of future trials of empathy training in healthcare can use the results of this review as a guide to their intervention development." Page 22
Conclusion 46-53 How would investigators use this review? The authors need to use this opportunity to spell out the raison d'etre of their study	We have taken note of this comment and have re-written the conclusion section: "Teaching students and other learners how to enhance empathy is moderately effective over a sustained period of time and is likely to benefit present and future patients. Future research should focus on empathy-interventions with patient-led outcome assessment and on assessing effectiveness of training over more sustained periods of time. Medical educators and curriculum designers can use this research to think of ways to integrate empathy training into busy curricula." Page 22

Table 1 The Gholamzadeh row: there is no entry in the effect of intervention column	This has been corrected. Page 11
Reviewer #2	
General Comments	
Thank you for the opportunity to review the article bmjopen-2019-036471 “Assessing the effect of empathy-enhancing interventions in health education and training: A systematic review of randomised controlled trials”. The subject is very relevant for the medical education field and offers a starting point for researchers interested in exploring the field. The overall quality of the article is good and some improvements can be easily achieved by the authors. I hope the commentaries below are helpful.	We thank the reviewer for their constructive comments which have been used to improve the manuscript.
Specific comments	
1 – In the introduction the authors state the importance of empathy for clinical care, they succinctly describe the core elements of empathy as a concept and share the previous attempts of reviewing the efforts to teach empathy. However, the discussion about how to measure empathy, and what empathy means in the clinical x non-clinical context is missing. Considering that the authors review focus on measuring empathy after specific interventions, the reader would benefit from being introduced to how empathy is or can be measured in the medical education field.	In response to this suggestion we have amended the introduction to acknowledge the difficulties with measuring empathy: “There is still however, little consensus on the precise nature of clinical empathy, not least reflected in the variety of tools and scales available to measure it. No guidance exists on how to select measures for assessing clinical empathy and choice of tools is likely to be led by the definition of empathy used or specific domain being measured.[11] A recent systematic review[11] on empathy measurement tools for care professionals identifies certain measures as scoring highest for quality, but concedes even these had low scores in some of the criteria they used.” Page 3-4

2 – Although the effort of authors in reviewing the literature and synthetizing the data is valuable, the articles are very heterogeneous, and we cannot be sure that these articles are measuring the same construct. So, I suggest the authors a separate analysis for the articles that are using the same kind of measurements, for instance CARE or JSE, which seems to be the most used, with the first focusing in self-assessment and the later on patient assessment.	We thank the reviewer for this comment and agree that the articles are heterogenous, which we expected and acknowledge as a limitation of the review. We have however further expanded on this in the limitations section, discussing specifically problems with the heterogeneity of measurement tools. We have acknowledged the heterogeneity of studies in our limitations but have now expanded on this to further acknowledge the differences in tools used to measure outcomes and that these may not all be measuring the same construct. "We recognise the heterogeneity of the studies in our review and anticipated this. This means that further research is required to identify the most effective empathy training methodology. Also, the strength of findings in this review may be limited by the reporting quality of some of the included studies. A sensitivity analysis of studies of highest quality found a slightly smaller but still significant effect size. Another limitation in reviewing the evidence in this field is the multiple tools used by investigators to measure clinical empathy. With the lack of a definitive definition of clinical empathy and a range of tools measuring different aspects of empathy, the impact of an intervention may vary depending on the measurement tool used. This is demonstrated by Reiss et al [44] who found a statistically significant improvement in empathy when measured using the CARE scale but no significant changes using the JSE. In contrast Buffel du Vaure [30] reported the opposite." Page 22 As a group we had considered whether to perform a separate analysis grouping studies by the measurement tool used, but instead opted to perform a sub-group analysis with self-assessed vs observer-assessed (eFigure 5), grouping studies this way rather than by individual measurement.
3 – The discussion starts with a statement that the interventions were efficient in improving empathy, but the authors are assuming that self-reported empathy is an accurate measurement of empathy, and	We have included an acknowledgement in the 'strengths and limitations' section of the paper to reflect the difficulties of using different tools to measure empathy. See above.

this is debatable. I think the authors should better elaborate on that, and without discussing what each one of the scales are measuring, it is impossible to be faithful to the findings and not misguide the reader.	
Reviewer #3	
General comments	
I have carefully reviewed the paper entitled « Assessing the effect of empathy-enhancing interventions in health education and training: A systematic review of randomised controlled trials. » submitted to Open BMJ by Rachel Winter et al. The present manuscript assess effect of empathy-enhancing interventions among student in the health field by performing a systematic review and meta-analysis. The manuscript was well written, the systematic review and meta-analysis was well done, following the PRISMA guideline with details on studies selection. As underlined by authors, a high heterogeneity was anticipated and observed likely due to the low quality of available RCT.	We thank the reviewer for their constructive comments, which has been used to improve the manuscript.
Specific comments	
The statistical section is well written, with a typo mistake in p-value to declare heterogeneity "(statistically significant for p0.01)"; I believe that authors wanted to write "heterogeneity was declared if p-value for heterogeneity<0.10 or I2>50%.	We agree with the reviewer and have amended this typo mistake. Page 9
There is some inconsistency between results section and flowchart (Figure 1) for the number of articles retrieved for full-text review (73 in the flowchart vs 72 in the text) and for the number of studies excluded (47 in the flowchart, 46 in the text). Please clarify.	We have reviewed this inconsistency and the data and have now ensured that Figure 1 represents the correct figures as represented in the full-text review (and eTable 2).

Similarly, there is some inconsistency between results section and others figures/tables: For example, in the result section you state that the duration of intervention ranges from 20 minutes to 18 hours, however in the Table 1, I found a duration of 42 hours for Yang's study.	We have amended the text to read 20 minutes to 42 hours having reviewed the data and thank the reviewer for drawing attention to this. Page 13
In the table 2, the heterogeneity for the study with least risk of bias is 66% whereas in the Figure 3 it is of 63%.	We have now corrected this inconsistency. Page 17-18
There is also inconsistencies between values of standardised mean difference and their 95% confidence interval in the Table 2 and the supplemental Figures for effect of duration of intervention and participant population.	We thank the reviewer for drawing our attention to these inconsistencies. We have revised Table 2 to reflect the data presented in the figures. Page 17-18
In the eTable 4, there is a typo mistake in the upper limit of the 95% CI of the risk with empathy training SMD, the value should be 0.67 instead of 0.37?	We have now corrected this error. eTable 4 Supplement

VERSION 2 – REVIEW

REVIEWER	Marco Antonio de Carvalho Filho Unversity Medical Center Groningen / University of Minho
REVIEW RETURNED	20-Jul-2020

GENERAL COMMENTS	The authors did a great job with their review and the article is ready for publication.

REVIEWER	Behal CHU Lille France
REVIEW RETURNED	23-Jul-2020

GENERAL COMMENTS	All questions and comments have been addressed.