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 Quality of Primary Health Care in 7 Chinese Provinces with Unannounced Standardized Patients: Protocol of a Cross-sectional Survey
AUTHORS	Xu, Dong; Hu, Mengyao; He, Wenjun; Liao, Jing; Cai, Yiyuan; Sylvia, Sean; Hanson, Kara; Chen, Yao-Long; Pan, Jay; Zhou, Zhongliang; Zhang, Nan; Tang, Chengxiang; Wang, Xiaohui; Rozelle, Scott; He, Hua; Wang, Hong; Chan, Gary; Melipillán, Edmundo; Zhou, Wei; Gong, Wenjie

VERSION 1 – REVIEW

REVIEWER	Sondra Zabar New York University School of Medicine, United States
REVIEW RETURNED	09-Jun-2018

GENERAL COMMENTS	This protocol paper was clearly written and covered all the key steps of delivering Unannounced Standardized Patient to assess quality of both clinical environment and health care delivered. There are few USP projects that look at systems as well as clinical performance. The team is experienced and realistic limitations are discussed in the protocol. Training includes field testing which is a key component to successful USP programs. Protocol address validity and reliability of SP performance as well as issue with first visit assessment and variation. Study does not mention how many per week and how they will achieve the almost 2000 visits they propose. There was no protocol for how detecting will be assessed and analysis of influence on outcomes of visit. There was not description of study team who will be delivering the USP across 7 provinces and if SP would be fielded in parallel or in series and or 10 cases will be distributed if each providence received approx 27 cases.

REVIEWER	Saul J. Weiner University of Illinois at Chicago and the Jesse Brown VA Medical Center, Chicago, IL USA
REVIEW RETURNED	17-Jun-2018

GENERAL COMMENTS	Unannounced standardized patient (USP) studies are complex and difficult to execute successfully. One of the investigators, Sean Sylvia, has considerable experience which is reassuring. There is Dr. Sylvia's study based on simulated cases of dysentery and angina, drawn from a prior project conducted in India by another team, and the more recent publication on tuberculosis detection in real China. The proposed study is much larger and more ambitious.

	Overall I believe the team has the background and a solid enough methodological framework to successfully carry out the study. However, I have a number of concerns about publishing this protocol in its current form. As indicated below there are problems with how it is organized and with missing elements. The concepts of measuring clinician performance and overall healthcare quality seem to be conflated. While the basic elements of a USP protocol are here, the specifics that will ultimately determine the quality of this study are not included. Specifically, we do not get to see any of the cases, or checklists, or scoring systems that will be used in the study. It is already evident from prior research what the general findings of the study will likely be: that healthcare quality as measured using unannounced standardized patients is very poor. The evidence for that is already overwhelming. So the question is, what will the real contribution of this study be? I think the answer to that lies in knowing precisely how and what they will measure. Without that information this is not a particularly innovative or informative protocol. Abstract The stated goal in the introduction section of the abstract of this protocol is “to collect quality information“ pertinent to primary health care in seven Chinese provinces. In the Methods and Analysis section it also says “several hypotheses will also be tested including the effect of facility ownership on PHC quality.” Proposed hypotheses should be explicitly stated in the introduction not the methods and analysis section. The methods section of the abstract says that “a standard protocol will be validated for validity…” that is circular reasoning. One has to specify the type of validity. It appears they are referring to construct validity. Also I think they mean they will look for evidence of construct validity, which is the correct terminology. They refer to doing “the usual descriptive analysis…“ I’m not sure what “usual“ refers to. Background Standardized patients are generally utilized to assess clinician performance, which is a component of quality but not the same as quality. Quality of care is an all encompassing term that takes into account all aspects of the care delivery system. Unannounced standardized patients have been considered the gold standard for performance assessment measurement, not quality measurement as the authors state. The rubric the authors have adapted, consisting of the six comprehensive means put forth by the IOM pertain to quality rather than performance. The authors appear to be making the case that USPs will be utilized for a comprehensive assessment of all six aims that comprise quality. It would be helpful for the authors to explain the rationale for applying this methodology beyond its usual assessment of performance to assess a global measure of quality. How, for instance, will USPs determine if lab tests are run correctly? That is an element of quality. Methods The authors refer to “creating a representative sample of China’s primary healthcare providers…“ and propose to do 1981 SP visits.

	Given that this is a descriptive study without any primary hypotheses driving the design, how do they calculate a sample size? In the section on sample size calculation they state that it was calculated “with the primary purpose of the standard descriptive survey analysis of this survey.“ I’m not able to follow the technical description in the rest of this paragraph because they reference materials and documents not available. Outcome variables Without actually having a sample case or checklist it is hard to assess this ambitious framework for globally measuring quality using the six IOM Aims. Note, I am not clear how “clinician politeness and friendliness“ are measures of “timeliness“. “a sperate” should be “separate." Other words are misspelled and there are grammatical errors throughout. Hypothesis Testing It appears that this is primarily a descriptive study with plans to do several hypothesis driven analyses, but they never explicitly state what the hypotheses are. Instead they say that they will “assess whether private providers provide inferior quality to public providers.“ Is their hypothesis that they will? Table 1: I notice that for five of the 10 conditions they include prescribing traditional Chinese drugs. Do they have research evidence to indicate if these are in fact evidence-based?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Sondra Zabar
Institution and Country: New York University School of Medicine, United States

Dr. Zabar: This protocol paper was clearly written and covered all the key steps of delivering Unannounced Standardized Patient to assess quality of both clinical environment and health care delivered. There are few USP projects that look at systems as well as clinical performance. The team is experienced and realistic limitations are discussed in the protocol. Training includes field testing which is a key component to successful USP programs. Protocol address validity and reliability of SP performance as well as issue with first visit assessment and variation.

Response: Many thanks for the encouragement from Dr. Zabar.

Dr. Zabar: Study does not mention how many per week and how they will achieve the almost 2000 visits they propose. There was not description of study team who will be delivering the USP across 7 provinces and if SP would be fielded in parallel or in series and or 10 cases will be distributed if each providence received approx 27 cases.

Response: We should have included more details in our original manuscript. We have added more information under several sections related to this question. Basically, we will recruit and train 7 SPs per case and assign cases to a random sequence of our 1981 SP-clinician visits. Then the 7 SPs per case will be dispatched concurrently to visit the facilities in accordance with the random sequence. Please see details of the revised text below:

"Each case will have 7 SPs who will be trained according to a standardized training manual that will be developed to guide the training and appraisal of the SPs." (Section "Selecting and Training SPs)

"After the facilities are selected and the number of visits per facility is determined, each of the planned visits will be given a unique identifier (e.g.: facility A-1, facility A-2, facility B-1), which will then be randomly ordered to form a random sequence numbered from 1 to 1981 consecutively. One of the 10 SP cases will be randomly assigned to each number on this random sequence. The 7 SPs per case will be dispatched to the 7 provinces concurrently, 1 SP per province. If multiple clinicians are available in that facility at the time of a particular SP visit (PHC visits in China do not require appointments), the field coordinator will randomly select a clinician by drawing lots onsite. Each SP is expected to make a total of approximately 30 visits. We plan to complete those SP visits over a three-month time span." (Section "Fielding the SPs)

Dr. Zabar: There was no protocol for how detecting will be assessed and analysis of influence on outcomes of visit.

Response: Thanks for this important comment. We have now added a new paragraph under the section "Fielding the SPs" to address this problem.

"In a separate but related study, a week after the visit of the SP, the same clinician will the same clinician will perform the same consultation but with a standardized virtual patient on a smartphone (Note: see https://bmjopen.bmj.com/content/8/7/e020943 for the published protocol for that study). We will use this opportunity to administer a detection questionnaire to the clinician, asking whether they suspect they had any visit from an SP over the past week. The detected cases will be treated as missing data in the data analysis." (Section "Fielding the SPs")

Reviewer: 2
Reviewer Name: Saul J. Weiner
Institution and Country: University of Illinois at Chicago and the Jesse Brown VA Medical Center, Chicago, IL USA

Dr. Weiner: Unannounced standardized patient (USP) studies are complex and difficult to execute successfully. One of the investigators, Sean Sylvia, has considerable experience which is reassuring. There is Dr. Sylvia's study based on simulated cases of dysentery and angina, drawn from a prior project conducted in India by another team, and the more recent publication on tuberculosis detection in real China. The proposed study is much larger and more ambitious.
Overall I believe the team has the background and a solid enough methodological framework to successfully carry out the study.

Response: We thank Dr. Weiner for recognizing our capability to conduct this study. Dr Sylvia has remained a core team member in this project. Meanwhile, Dr. Zhongliang Zhou, another core member of our team, has successfully completed quality studies in Shan'xi province, using 3 SP cases. . Although study results have not been published, we have been able to accumulate considerable experience in designing and implementing USPs.

Dr. Weiner: However, I have a number of concerns about publishing this protocol in its current form. As indicated below there are problems with how it is organized and with missing elements. The concepts of measuring clinician performance and overall healthcare quality seem to be conflated. While the basic elements of a USP protocol are here, the specifics that will ultimately determine the quality of this study are not included. Specifically, we do not get to see any of the cases, or checklists, or scoring systems that will be used in the study. It is already evident from prior research what the

general findings of the study will likely be: that healthcare quality as measured using unannounced standardized patients is very poor. The evidence for that is already overwhelming. So the question is, what will the real contribution of this study be? I think the answer to that lies in knowing precisely how and what they will measure. Without that information this is not a particularly innovative or informative protocol.

Response: In this protocol, we have tried to provide as many and specific details as possible subject to the constraint of producing a readable paper length. More details have been added to the revised manuscript. We understand your concerns about the conflation of clinician performance and overall quality of care. We will return to this in our response to your following comments.

As for the requested specific cases, or checklists, we regret that we will not be able to provide those documents at this time of our study. Please note that this manuscript is a study protocol which intends to outline the methods to design, develop, and implement this study using USPs. We have not yet completed the development of our cases so we are unable to provide samples. Although we have some prototypes of the cases, we are reluctant to release these as none of them has gone through the validation and revision process we have outlined in this study protocol.

We understand that without the provision of those specific cases and considering the general finding of poor clinician performance as assessed by the USP, it may be difficult to evaluate the innovation and usefulness of this study. Yet, we would like to mention the strong policy and research relevance of this study. Policy-wise, there have been few studies of quality of primary health care in China - particularly those using a representative sample and using a rigorous method of assessment. Even though we expect our study to detect poor quality in line with prior studies, we still need to know precisely the quality information that our diverse group SP cases will be able to generate. In particular, we selected the specific SP cases on the basis of evidence about the distribution of common conditions in China's PHC setting. With these cases, we will be able to map out a more comprehensive picture of the quality in China's PHC. In terms of novel research findins, this study will produce the basic data that will be used in a series of hypothesis-driven studies. As you have pointed out, we should have provided more details on those hypothesis-driven studies. We address this issue in our response to your following questions.

Dr. Weiner: The stated goal in the introduction section of the abstract of this protocol is "to collect quality information" pertinent to primary health care in seven Chinese provinces. In the Methods and Analysis section it also says "several hypotheses will also be tested including the effect of facility ownership on PHC quality." Proposed hypotheses should be explicitly stated in the introduction not the methods and analysis section.

Response: Thanks for this important input and the opportunity for us to clarify this issue. This study is the beginning of a series of studies we have planned. This first one will primarily be a descriptive study based on the USP survey data. The purpose is to present the quality of PHC in China. Thus, we feel it is better not to list our hypotheses in the introduction section as it may confuse readers on the main purpose of this study protocol (ie., descriptive rather than hypothesis-testing in nature). This first study will set foundation for our other related studies. The details of those other hypothesis-driven studies will be described in separate study protocols, some of which have been published (https://bmjopen.bmj.com/content/8/7/e020943). We have now added a statement at the end of the "Introduction" section and have fully revised the section "hypothesis-driven analysis" as "Related Studies":

"The project has involved 20 universities across 19 provinces in China as well as researchers from Nepal, USA, and UK in a USP Network (https://www.researchgate.net/project/Unannounced-

Standardized-Patient-USP-and-Virtual-Patient-VP-to-Measure-Quality-of-Primary-Care). The USP resources will be pooled and shared widely within the network first and then with the general public. This study is the first of a series of studies to be based on the quality data collected using USPs. The primary purpose of this study is to collect and present descriptive data on the quality of China's PHC. We have developed / are developing separate protocols for the various hypothesis-driven studies, which will be available elsewhere and from our Network website.29 (Section "Introduction")

"This study protocol mainly deals with the descriptive analysis and presentation of the data to be collected by the USPs. Using the USP survey data, we have planned several related studies that will be covered by separate study protocols with detail on the background, theoretical framework, and analytical methods. To summarize those related studies, we will assess (1) the effect of ownership types of the PHC providers (i.e., private versus public) on the quality of PHC (study protocol under revision), (2) the know-do gap between the assessment results by a smartphone-based virtual standardized patient and USP (protocol already published),29 (3) the effect of using smartphone-based virtual patient in improving clinician performance, (4) the effect of type of insurance on the quality of care, (5) the impact of gatekeeping by primary care providers on the quality of TB care – a mathematical modeling study, and (6) clinician skills in handling low-value or harmful patient requested services – particularly antibiotics and some processed traditional Chinese medicine." (Section "Related Studies")

Dr. Weiner: The methods section of the abstract says that "a standard protocol will be validated for validity…" that is circular reasoning. One has to specify the type of validity. It appears they are referring to construct validity. Also I think they mean they will look for evidence of construct validity, which is the correct terminology.

Response: Thanks for pointing out this ambiguity. By "will be validated for validity and reliability", we mean that the SP cases will be validated before use, and the validation will address several kinds of validity and reliability as well. We have now revised the abstract to make it clearer. We have also revised our "USP Validation" part at the method section to provide more details on our validation approach along with a table that summarizes our validation methods.

"The SP cases and the checklist will be developed through a standard protocol and will be assessed for content, face and criterion validity and test-retest and inter-rater reliability before its full use." (Section "Abstract")

"USP validation will be based on a convenience sample of clinicians not included in our final survey sample in the project training and pilot phase. Those SP-clinician interactions in the pilot will be audio recorded and transcribed. Validity is the extent to which an instrument measures what it is supposed to measure. We will assess content, face, and criterion validity of the cases. The content validity will be assessed by an expert panel who will use a 4-point Likert scale to evaluate the appropriateness of the written content of the cases. The face validity of the SP assessment depends on (1) SP remaining undetected (detection ratio reported to be 5%-10%55), and (2) authentically and consistently portraying the clinical features of the case. We will send the participating clinician in the pilot a "detection form" to report their degree of suspicion of any SP visit.46 The authenticity of the SP presentation will be evaluated by checking the transcribed recording whether a key piece of information was divulged by the SP when appropriately prompted, not divulged when prompted, or volunteered when not prompted. Criterion validity will be assessed through the agreement of the SP-completed checklist against that completed by a clinician based on the transcript of the visit (i.e., the clinician rating as the "gold standard").56-59 Checklist items which depend on visual observation will be excluded. Reliability examines the level of consistency of the repeated measurements. The inter-rater reliability of two SPs on the same condition and context will be assessed with two SPs completing the checklist for the same recorded transcript. Test-retest reliability will be analyzed by the

concordance of assessment results of the same SP to score his or her own recorded encounter a month later).57 The agreement will be analyzed with Lin's concordance correlation coefficient (rc)60. rc indicates how closely pairs of observation fell on a 45° line (the perfect concordance line) through the origin in addition to their correlation.60-62 Bland-Altman plot will be used to visualize the concordance.63 64 Table 3 summarizes our methods of validation." (Section "USP Validation")

Dr. Weiner: They refer to doing "the usual descriptive analysis…" I'm not sure what "usual" refers to.

Response: Thank you for pointing this out. We agree that the use of the term "usual" is not clear and have now removed this from our abstract. Furthermore, we added more details to the descriptive analysis we would perform at the method section:

"We will focus on descriptive analysis to present quality of PHC in those 7 provinces. Hypothesis-driven analyses will be described in separate study protocols. For descriptive analysis, we will first present clinician and facility profiles in tables for all 7 provinces and by each province. The clinician profile will include socio-demographic information (age, gender, and ethnicity), professional qualification (general and medical education, licensure, and professional ranks), and service information (volume of visits, number of support personnel). The Facility profile will include operation and management (years in operation, ownership types, accreditation, level of hospitals, affiliation with medical universities, revenue, health insurance contracting, payment methods), clinical services (annual number of inpatient and outpatient visits, number of clinical departments), personnel (number of physicians, nurses, and attrition ratio), and equipment. Secondly, we will tabulate results of overall quality and sub-domains across administrative regions and provider types. Thirdly, we will map out the locations of the facilities along with their quality scores with geospatial analytical tools. Finally, T-test/Wilcoxon test or Chi-square test will be employed to compare quality differences of public versus private providers, primary care clinics/centers versus hospital outpatient services, and rural versus urban areas, and across different conditions, clinician educational levels, and payment mechanisms." (Section "Survey Analysis")

Dr. Weiner:
Background
Standardized patients are generally utilized to assess clinician performance, which is a component of quality but not the same as quality. Quality of care is an all encompassing term that takes into account all aspects of the care delivery system. Unannounced standardized patients have been considered the gold standard for performance assessment measurement, not quality measurement as the authors state. The rubric the authors have adapted, consisting of the six comprehensive means put forth by the IOM pertain to quality rather than performance. The authors appear to be making the case that USPs will be utilized for a comprehensive assessment of all six aims that comprise quality. It would be helpful for the authors to explain the rationale for applying this methodology beyond its usual assessment of performance to assess a global measure of quality. How, for instance, will USPs determine if lab tests are run correctly? That is an element of quality.

Response: Many thanks for explaining the important distinction between clinician performance and overall quality. We agree that USP is the gold standard for measuring clinician performance but not the overall quality. However, we have aldo tried to introduce the additional elements of quality from the IOM quality framework for which the SP can evaluate or collect information for later evaluation. However, we agree with you that even after the adoption of this IOM framework of comprehensive evaluation, the evaluation is still largely around clinician performance. It could be argued that the critical determinant of quality in the primary health care setting is clinician performance. Nonetheless,

we feel it is appropriate to state this as a limitation of the study and have thus revised our discussion section to reflect this limitation.

“The study has several potential limitations. First of all, even though the assessment of SP is considered the gold standard for measuring clinician performance, and in this study we have further expanded the use of SPs to evaluate other elements of quality in the IOM framework like patient-centeredness, timeliness, and efficiency, we should recognize that all those quality elements are still largely clinician-related, and other important quality aspects, such as the quality of laboratory testing, cannot be assessed by our SPs.” (Section “Discussion)

Dr. Weiner:
Methods
The authors refer to “creating a representative sample of China’s primary healthcare providers…“ and propose to do 1981 SP visits. Given that this is a descriptive study without any primary hypotheses driving the design, how do they calculate a sample size? In the section on sample size calculation they state that it was calculated “with the primary purpose of the standard descriptive survey analysis of this survey.“ I’m not able to follow the technical description in the rest of this paragraph because they reference materials and documents not available.

Response: As this is primarily a descriptive study, we follow the method of the sample size calculation for descriptive studies based on surveys rather than the usual method for hypothesis-driven studies. Survey sample size was calculated to achieve the desired level of relative precision (coefficient of variation, CV) of the variable of the interest rather than to detect a certain level of effect size.

Dr. Weiner:
Outcome variables
Without actually having a sample case or checklist it is hard to assess this ambitious framework for globally measuring quality using the six IOM Aims. Note, I am not clear how “clinician politeness and friendliness“ are measures of “timeliness“.

Response: Again, we understand that it is difficult to assess our framework of using SP to assess overall quality and we apologize for not being able to provide a sample case and checklist as all the cases are under development. Our first paper (not this protocol paper) will present theresults of our case validation that will include both sample cases and checklists. We also agree with Dr. Weiner that “clinician politeness and friendliness” are not measures of “timeliness”. We have removed this statement from the manuscript. Note that clinician politeness and friendliness have already been covered by the Patient Perception of Patient-centeredness (PPPC) rating scale

Dr. Weiner: “a sperate” should be “separate." Other words are misspelled and there are grammatical errors throughout.

Response: We are sorry for those typos and grammatical errors. We have now given a thorough check of the manuscript for errors.

Dr. Weiner:
Hypothesis Testing
It appears that this is primarily a descriptive study with plans to do several hypothesis driven analyses, but they never explicitly state what the hypotheses are. Instead they say that they will “assess whether private providers provide inferior quality to public providers.“ Is their hypothesis that they will?

Response: Yes, this is primarily a descriptive study. We will conduct several hypothesis-driven studies that will need separate study protocols. However, taking the advice from Dr. Weiner, we have revised our manuscript concerning the section of “hypothesis Testing (section title now changed to be “Related Studies”” to give a summary of our planned related studies:

“This study protocol mainly deals with the descriptive analysis and presentation of the data to be collected by the USPs. Using the USP survey data, we have planned several related studies that will be covered by separate study protocols with detail on the background, theoretical framework, and analytical methods. To summarize those related studies, we will assess (1) the effect of ownership types of the PHC providers (i.e., private versus public) on quality of PHC (study protocol under revision), (2) the know-do gap between the assessment results by a smartphone-based virtual standardized patient and USP (protocol already published),29 (3) the effect of using smartphone-based virtual patient in improving clinician performance, (4) the effect of insurance types of a patient on quality of care, (5) the impact of gatekeeping by primary care providers on quality of TB care – a mathematical modeling study, and (6) clinician skills in handling low-value or harmful patient requested services – particularly antibiotics and some processed traditional Chinese medicine.” (Section “Related Studies”)

Dr. Weiner:
Table 1: I notice that for five of the 10 conditions they include prescribing traditional Chinese drugs. Do they have research evidence to indicate if these are in fact evidence-based?

Response: Thanks, Dr. Weiner, for this opportunity to clarify what we mean. In the Table, we include some cases that will specifically deal with the issue of the use of processed traditional Chinese medicines (TCM). We are not saying the use of those medicines is evidence-based - on the contrary, many of them are not evidence-based. The cases in which TCM are indicated are the cases which other studies or our observations have shown that TCM is commonly used (perhaps inappropriately). One purpose of our study is to use those cases to track how exactly how TCM is used by primary care providers (most of whom are western rather than TCM doctors).

VERSION 2 – REVIEW

REVIEWER	Saul J. Weiner University of Illinois at Chicago College of Medicine
REVIEW RETURNED	12-Sep-2018

GENERAL COMMENTS	The authors have done a nice job of addressing various concerns. I do not have prior experience reviewing and recommending for publication protocols of studies not yet undertaken. However, this one if it is carried out, will be particularly significant because it will be a nationwide assessment of healthcare quality based on direct covert observation, which is novel and considered by some to be a “gold standard measure” of physician performance (In fact the “gold standard” rating for USP was made in a BMJ publication on the topic over a decade ago). Page 3. “The seven provinces are not randomly selected, although we intend them to represent different health development conditions of China’s provinces.” What is a “health development condition”?

	P4: “we take the IOM…” should be “we adopt…”; “clinical guideline” should be “clinical guidelines” P6: “licensed physician and licensed assistant physician” should both be plural P15: “Content validity will be assessed by an expert panel who will use a 4-point Likert scale to evaluate the appropriateness of the written content of the cases.” Actually, it’s critical that the cases and physician rating instruments are approved not only by an expert panel, but are based on evidence-based practices and top quality published guidelines. For instance, if a case is designed to assess physician performance at managing a patient with depression, or diabetes, or asthma etc... the checklists that standardized patients use should be developed so that they assess adherence to guidelines. This is really important! Otherwise, the results of the study will be challenged by those who say the “expert panel“ got it wrong. P 26: The proposal doesn’t provide any information on the funds allocated to this very ambitious and costly project. I think it is prudent to assure that the funds allocated are adequate to carry out this study.

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 2

Reviewer Name: Saul J. Weiner

Institution and Country: University of Illinois at Chicago

Please state any competing interests or state ‘None declared’: None

Please leave your comments for the authors below

The authors have done a nice job of addressing various concerns. I do not have prior experience reviewing and recommending for publication protocols of studies not yet undertaken. However, this one if it is carried out, will be particularly significant because it will be a nationwide assessment of healthcare quality based on direct covert observation, which is novel and considered by some to be a “gold standard measure” of physician performance (In fact the

“gold standard” rating for USP was made in a BMJ publication on the topic over a decade ago).

Response: Thank you, Dr. Weiner, for this overall comment.

Page 3. “The seven provinces are not randomly selected, although we intend them to represent different health development conditions of China’s provinces.”

What is a “health development condition”?

Response: We used the life expectancy of each province as the proxy for the health development condition of those provinces. We have now revised our manuscript to clarify this point:

“The seven provinces are not randomly selected, although we intend them to represent different health development conditions (by using life expectancy as the proxy) of China’s provinces.

P4: “we take the IOM…” should be “we adopt…”; “clinical guideline” should be “clinical guidelines”

Response: Revised accordingly.

P6: “licensed physician and licensed assistant physician” should both be plural

Response: Revised accordingly.

P15: “Content validity will be assessed by an expert panel who will use a 4-point Likert scale to evaluate the appropriateness of the written content of the cases.”

Actually, it’s critical that the cases and physician rating instruments are approved not only by an expert panel, but are based on evidence-based practices and top quality published guidelines. For instance, if a case is designed to assess physician performance at managing a patient with depression, or diabetes, or asthma, etc... the checklists that standardized patients use should be developed so that they assess adherence to guidelines. This is really important! Otherwise, the results of the study will be challenged by those who say the “expert panel“ got it wrong.

Response: We completely agree with Dr. Weiner on this point with regard to the checklist (i.e. the quality criteria). We should clarify that when the expert panel is reviewing the cases, they will be instructed to review the checklist in accordance with the quality clinical guidelines (which we will provide to them). However, they will also be reviewing other parts of the case such as scenario and script to check whether they actually reflect the clinical situation in accordance with their clinical experiences. Therefore, the expert panel is still necessary. We have revised this part as below in the manuscript:

“Content validity will be assessed by an expert panel who will use a 4-point Likert scale to evaluate the appropriateness of the written contents of the cases that will include the scenario, scripts, and checklists. For the checklist, they will be instructed to check the appropriateness against the published clinical guidelines.”

P 26: The proposal doesn't provide any information on the funds allocated to this very ambitious and costly project. I think it is prudent to assure that the funds allocated are adequate to carry out this study.

Response: In the original manuscript, there is a section after the tables at the end called "Funding" where we listed the grants that are supporting this study.

VERSION 3 – REVIEW

REVIEWER	Saul Weiner University of Illinois at Chicago
REVIEW RETURNED	04-Nov-2018

GENERAL COMMENTS	Looks good (this is my third review of this manuscript). Of note, in my prior review I asked for evidence the project is adequately funded as it seems large and costly, covering 7 provinces. The authors responded by noting they included the names of 3 grants. They didn't indicate the size of them though. I defer to the journal editors regarding whether they need this level of detail.