PEER REVIEW HISTORY

BMJ Paediatrics Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Paediatrics Open. The paper was subsequently accepted for publication at BMJ Paediatrics Open.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Poor inter-observer agreement in the measurement of respiratory rate in children: a prospective observational study
AUTHORS	Daw, William; Kingshott, Ruth; Elphick, Heather

VERSION 1 - REVIEW

REVIEWER	Brown, Nick Salisbury District Hospital, Paediatric department Competing Interests: None
REVIEW RETURNED	19-Mar-2017

GENERAL COMMENTS	This is an important, if not entirely novel area Though the analyses seem appropriate, I have some comments and concerns about some aspects of the design and interpretation 1. The implication here (rightly or otherwise) is that the first assessors (RR1, mainly nurses) overestimated RR possibly, in part, because they measured only for 15 seconds and then multiplied. This is perhaps not surprising if one considers that each extra included breath every 15 seconds (and an observer would naturally round up) is in effect, 4 for each minute which is approximately the degree of mean bias seen. This, I think should be the main message 2. The RR1 measures however, were not made simultaneously with the RR2 and 3s and we can't say for sure that the rates were not, in fact, genuinely different. It is entirely possible that the anxiety around admission (when the RR1 measures were made) raised the observed RRs and that the differences were entirely genuine rather than the result of measurement error. You say the time lag was 'far superior' to other studies in the area, but, I cannot understand why, if RR2 and 3 could be made at the same time, that all three measures could not have been simultaneous. You need to be more critical of this limitation 3. I don't agree with your comment that tachypnoeic children were missed as your data shows the opposite with the RR1 rates being consistently higher. In other words, more children were classified as tachypnoeic, at type 1 rather than type 2 error General: (a) Please avoid comments like 'interestingly' and 'superior to other studies'. You might think so, but readers have to decide for

	themselves (b) Detail. Rather too much and the messages get lost in the verbiage. Did having a third observer clear or muddy the picture ? (c) sample size. You have not referenced the 'previous study' on which your estimate is based. Why did you choose a difference of 2 breaths/minute ? Was your estimate bidirectional ? (d) Decimal places. In some places you have used 3 decimal places which gives the data an air of specious precision. See Tim Cole's piece for guidnce (http://dx.doi.org/10.1136/archdischild-2014-307149)

REVIEWER	Simoes, Eric Children´s Hospital Colorado, Paediatric Competing Interests: None
REVIEW RETURNED	18-May-2017

GENERAL COMMENTS	Major comments: For this study to be useful, the authors need to provide several more levels of detail that are not provided. These will be outlined in the detailed review below. A second major issue is trying to make comparisons with WHO cutpoints. These are actually completely wrong. The IMCI reference [reference 9] only refers to children under the age of five. Yet the authors seem to make inferences in table 1 about children greater than 60 months of age. These are not validated and are not part of the WHO guidelines. Even for those less than 60 months [even the age range given are wrong] the cutpoints of >60, > 50, > 40 are also wrong. The authors should carefully review the actual guidelines and make correct cutpoints for these age groups. it has been very difficult to define pneumonia in older ages and the WHO does not attempt to do this. Detailed comments: Introduction: 1. The main justification for the study appears to be that there are no studies in the UK determining the degree of interobserver agreement and respiratory rate measurements in children. As the authors themselves have quoted, there are many studies that have looked at interobserver agreement on respiratory rate in the Americas (Ref 20 of manuscript) And in developing countries summarized in [PLoS ONE 11(3): e0152204 doi:10.1371/journal.pone.0152204] 2. it is also not true that there are no reliable electronic devices that are available for measurement of the respiratory rate. The authors did not use one, but the justification is not that there are no reliable devices. Methods: 1. It is well-known that various states of alertness and agitation and fever affect the respiratory rate as well as its variability [Ref 10 of manuscript]. Clearly if children were recruited from all over the hospital, these simple parameters should have been recorded at least when the two study staff [RR2 and RR3] examine the patient. Thus it is clear that patients with bonds, head injuries, seizures, pain, diarrhea and vomiting for example, will have very variable for less than 30 seconds. Including this vast array of subjects is not

	helpful for the study and in fact completely influences the outcomes. 2. If these states have been collected, as they should have they should be presented at the analysis done taking this into account. 3. Table 1 should be corrected after careful perusal of reference 9. Results: 1. While the age range is wide for children studied, there is no breakdown of the ages so that one can determine the usefulness. 2. It is well-known that counting the respiratory rate for less than 60 seconds is quite inaccurate. It is not surprising that the nurses [R1] have the most variability, and with the least accuracy. Since most of the poor correlation refers to the R1 – R2 and R1 – R3 interobserver variability, this is not surprising at all and this does not had a call to the literature. All it tells us is that those hunting methods are poor. 3. There is no description of how the results in table 4 were obtained 4. Assessment of tachypnea: for starters, the WHO algorithm should only be applied to children less than 59 months of age with cough or difficult breathing. Since most of these children do not have cough or difficult breathing [114 of the 169] and it is unclear which of the 55 with cough or difficult breathing are less than 59 months of age, the whole analysis of tachypnea means nothing. Discussion: 1. Please see comment number one in the introduction relating to the first paragraph 2. on page 10, the authors refer to changes in the activity status of the child between measurements, not having an impact on the agreement. Perhaps this reflects that they did collect activity information. The data should be presented and analysis done by different activity states. 3. Frankly all the R1 – R2 and R1 – R3 comparisons, only a reflection of the poor method for collection by the nurses and this should be emphasized rather than the comparisons. using a poor method of data collection does not justify the title “Poor interobserver agreement and the measurement of restoration children” in fact the interobserver agreement for the R2 – R3 comparisons was reasonable as those shown in several of the studies referred to above. In fact the agreement is quite good. 4. The whole discussion about tachypnea should be revisited. Any comparison with the WHO guidelines should be carefully re-examined. 5. Most of the conclusions that the authors draw have been shown in several other studies over the years, but perhaps the most important observation that they have, is that if nurses are going to count the respiratory rate, they should be trained in doing it properly or not at all.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1 (Comments to the Author):

This is an important, if not entirely novel area.
Though the analyses seem appropriate, I have some comments and concerns about some aspects of the design and interpretation.

We thank the reviewer for their comments and have responded to each of the comments below.

1. The implication here (rightly or otherwise) is that the first assessors (RR1, mainly nurses) overestimated RR possibly, in part, because they measured only for 15 seconds and then multiplied. This is perhaps not surprising if one considers that each extra included breath every 15 seconds (and an observer would naturally round up) is in effect, 4 for each minute which is approximately the degree of mean bias seen. This, I think should be the main message

    We thank you for this comment and have emphasised this point within the discussion section referencing similar findings from previous studies. (Discussion paragraph 4)

2. The RR1 measures however, were not made simultaneously with the RR2 and 3s and we can't say for sure that the rates were not, in fact, genuinely different. It is entirely possible that the anxiety around admission (when the RR1 measures were made) raised the observed RRs and that the differences were entirely genuine rather than the result of measurement error. You say the time lag was 'far superior' to other studies in the area, but, I cannot understand why, if RR2 and 3 could be made at the same time, that all three measures could not have been simultaneous. You need to be more critical of this limitation

Thank you, this has raised an important point. The respiratory rate measurements taken in this study were never made upon admission. In fact all children had had more than one previous measurement made prior to all of our study measurements and they were all clinically stable. This would hopefully have negated any possible anxiety that they would have had around the admission period that may have falsely altered their RR. We have clarified this in paragraph 2 of the Methods section.

We agree that RR2 and RR3 could have been taken simultaneously with RR1, however we opted to take this after the initial HCP measurement so that we could also look at their actual clinical practice. We feel if the HCP had been aware of us taking the RR simultaneously with them then this would have possibly altered their method of measurement and would not have truly reflected their actual practice. This point was discussed in detail with our research ethics committee as we did not want to introduce this potential bias of HCPs being aware that their RR measurements were being observed by researchers. It was agreed that ward staff were made aware in advance that a research study was in progress, but individual HCPs were not aware that individual measurements were being recorded and repeated.

We feel that compared to other similar studies the time we used between measurements is equal to and in some cases a lot less. We have cited these studies below along with the times they left between measurements.

We have changed the comment about our study being 'far superior' in paragraph 5 of the Discussion.

Citation	Time between each measurement
Chan et al. Interobserver variability of croup scoring in clinical practice. Paediatric Child Health. 2001	- Within 1 hour
Wang et al. Observer agreement for respiratory signs and oximetry in infants hospitalised with lower resp infections. Am Rev Respir Dis. 1992	- Within 30 minutes

Citation	Time between each measurement
Wang et al. Study of observer reliability in clinical assessment of RSV lower respiratory illness (PICNIC). Paediatric Pulmol. 1996.	-Mean time between measurement = 90 mins. Some measurements up to 6 hours later
Liu et al. Use of a respiratory clinical score among different providers. Pediatr Pulmonol. 2004.	- No details given
Gajdos et al. Inter-observer agreement between physicians, nurses and respiratory therapists for respiratory clinical evaluation of bronchiolitis. Pediatr Pulmonol. 2009.	- Minimum of 8hrs between each assessment
Lanaspa et al. High reliability in respiratory rate assessment in children with resp symptomatology in a rural area in Mozambique. J Trop Pediatr. 2014	- Measurements taken with 30 minutes

3. I don't agree with your comment that tachypnoeic children were missed as your data shows the opposite with the RR1 rates being consistently higher. In other words, more children were classified as tachypnoeic, at type 1 rather than type 2 error

We have been through our data and for all those children with respiratory rates in the "tachypnoea" range, ie at or above the 95th centile for the child's age, the measurement taken by RR1 was higher in 63% of measurements. In 28% (15 children) RR1 did not classify the child as having a raised RR but one of or both of the other observers did. We have included this data within paragraph 8 of the Results section, but please also note we have now changed how we are defining tachypnoea (see reviewer 2's comments, below).

General:

(a) Please avoid comments like 'interestingly' and 'superior to other studies'. You might think so, but readers have to decide for themselves

We have removed the comment 'far superior' from the discussion section as noted above and also changed the first line of paragraph 6 of the Discussion.

(b) Detail. Rather too much and the messages get lost in the verbiage. Did having a third observer clear or muddy the picture?

We have altered sections of the discussion to include actual figures rather than descriptions and opinion. We feel that the third observer added to the study by giving a simultaneous RR measurement with which comparisons could be drawn. As explained above we did not feel it was right to complete a simultaneous measurement at the time the HCP took the measurement therefore in order to assess the agreement between simultaneous measurements we required a third observer.

(c) sample size. You have not referenced the 'previous study' on which your estimate is based. Why did you choose a difference of 2 breaths/minute ? Was your estimate bidirectional ?

Thank you for this comment, we have since referenced this study (and we are happy to provide this data upon request). The 2 breaths/min was bidirectional and was based on previous reported limits of agreement by the same observers in adults (Reference 17 - Liu, L.L., et al., Use of a respiratory clinical score among different providers. Pediatr Pulmonol, 2004. 37(3): p. 243-8.). The limits of agreement in this study were 5 breaths per minute but we wanted to select a larger sample size that would be able to detect a narrower range of limits of agreement than this and so ± 2 breaths/min was selected.

(d) Decimal places. In some places you have used 3 decimal places which gives the data an air of specious precision. See Tim Cole's piece for guidance (http://dx.doi.org/10.1136/archdischild-2014-307149)

We appreciate the guidance with regards this and have since changed our values to 1 decimal place.

Reviewer 2 (Comments to the Author):
Major comments
For this study to be useful, the authors need to provide several more levels of detail that are not provided. These will be outlined in the detailed review below.

We thank the reviewer for their comments and we will address these in order below.

A second major issue is trying to make comparisons with WHO cutpoints. These are actually completely wrong. The IMCI reference [reference 9] only refers to children under the age of five. Yet the authors seem to make inferences in table 1 about children greater than 60 months of age. These are not validated and are not part of the WHO guidelines. Even for those less than 60 months [even the age range given are wrong] the cutpoints of >60, > 50, > 40 are also wrong. The authors should carefully review the actual guidelines and make correct cutpoints for these age groups. it has been very difficult to define pneumonia in older ages and the WHO does not attempt to do this.

Thank you for drawing this to our attention. We had taken the cut off points for those children over 5 years from a different source and indeed these are not part of the WHO guidelines. We have omitted the references for these age groups and this was an oversight, for which we apologise. In reviewing the reference ranges used here we have decided to opt instead for respiratory rates that were ≥95th centile as defined by APLS guidelines. As there is still a growing body of evidence as to what constitutes a normal respiratory rate we feel by using this cut off point it will avoid confusion and hopefully be of greater value for readers to be able to interpret our data within a clinical context.

We have altered Table 1 to reflect these new values and subsequent results and discussion are written based upon these values.

Detailed comments:
Introduction:
1. The main justification for the study appears to be that there are no studies in the UK determining the degree of interobserver agreement and respiratory rate measurements in children. As the authors themselves have quoted, there are many studies that have looked at interobserver agreement on

respiratory rate in the Americas (Ref 20 of manuscript) And in developing countries summarized in [PLoS ONE 11(3): e0152204 doi:10.1371/journal.pone.0152204]

> *We feel that although there are multiple other studies assessing the inter-observer agreement of RR measurements, these studies are very heterogeneous and report a wide range of variability, mainly in the form of an intraclass correlation coefficient (including reference 20 - [PLoS ONE 11(3): e0152204 doi:10.1371/journal.pone.0152204) Three of the studies only looked at a small convenience sample, two only looked at specific illnesses or children within a very narrow age range and the two larger studies compared respiratory rates taken on average 90 minutes apart. We have attempted to produce a study that could address this issue and bring a more conclusive answer. We have included a table below outlining the current evidence in this area which highlights the inconsistencies of the current available evidence.*

Studies assessing inter-observer variability in the measurement of respiratory rate in children

Citation	Study Group	Study Type	Methods	Relevant Key Results	Comments
Chan et al. Interobserver variability of croup scoring in clinical practice. Paediatric Child Health. 2001	158 Children aged 3 months - 5 years presenting with viral croup	Prospective cohort study	Child assessed by triage nurse, ED nurse and ED physician within 1 hour for clinical signs associated with croup - including RR	Weighted Kappa score for RR agreement: Traige nurse v ED nurse: 0.17 ED Nurse v Physician: 0.15 Traige nurse v ED Physician: 0.24	- Only accounts for children presenting with viral croup - 1 hr window may lead to variation in clinical status. - RR converted to categorical score - Large cohort studied - RR counted over 30 seconds then doubled
Wang et al. Observer agreement for respiratory signs and oximetry in infants hospitalised with lower resp infections. Am Rev Respir Dis. 1992	56 infants <2yrs hospitalised with bronchiolitis or pneumonia	Prospective cohort study	Assessed by Paediatric infectious disease consultant + Infectious disease nurse or infectious disease fellow. RR measured within 20 minutes	Kappa score for RR agreement: 0.38	- Small convenience sample - RR counted over 30 seconds - -RR converted to categorical score - Study ran over two 3 month periods 2 years apart

Studies assessing inter-observer variability in the measurement of respiratory rate in children

Citation	Study Group	Study Type	Methods	Relevant Key Results	Comments
Wang et al. Study of observer reliability in clinical assessment of RSV lower respiratory illness (PICNIC). Paediatric Pulmol. 1996.	137 infants with RSV respiratory illness across 8 centres	Prospective cohort study	Two blinded observers: Research nurse + nurse or Paediatrician	Pearson correlation coefficient for RR agreement = 0.42 - 0.97	-RR counted over a full minute -Some assessments took place 6 hrs later with mean = 90 mins -Highest agreement seen in centre with fewest recruits
Liu et al. Use of a respiratory clinical score among different providers. Pediatr Pulmonol. 2004.	55 patients <1yr-19yrs admitted with asthma bronchiolitis or wheezing	Prospective cohort study	Physicians, nurses and respiratory therapists simultaneously assessed RR	Kappa score (unweighted) 0.36 (95% CI 0.26-0.46)	- Small convenience sample - RR converted to categorical score - No details of how RR measured given - Large age range of children studied
Gajdos et al. Inter-observer agreement between physicians, nurses and respiratory therapists for respiratory clinical evaluation of bronchiolitis. Pediatr Pulmonol. 2009.	180 infants under 18 months hospitalised with 1st episode of bronchiolitis	Prospective cohort study	Physicians, nurses and respiratory therapists. Two providers assessed child's RR at same time	Weighted Kappa score : 0.76 - 0.97. Highest agreement seen between 2 physicians	- Only accounts for infants with bronchiolitis - Narrow age range of children studied - No details of how RR measured - Minimum of 8hrs between each assessment - RR converted to categorical score

Studies assessing inter-observer variability in the measurement of respiratory rate in children

Citation	Study Group	Study Type	Methods	Relevant Key Results	Comments
Lanaspa et al. High reliability in respiratory rate assessment in children with resp symptomatology in a rural area in Mozambique. J Trop Pediatr. 2014	55 children <10 years with cough, fever, or breathing difficulties in developing country setting	Prospective cohort study	RR measured 3 times by different observers in 30 min period.	Agreement in RR count Intraclass Correlation Coefficient of 0.95 (95% CI: 0.93-0.97).	- RR counted over 60 seconds - Observers - medical agent + 2 study health assistants - Small sample size - Children from developing country

2. it is also not true that there are no reliable electronic devices that are available for measurement of the respiratory rate. The authors did not use one, but the justification is not that there are no reliable devices.

We apologise for any confusion created here. We were wanting to convey that although there are devices used to monitor children's respiratory rate there are no devices that exist to provide a rapid one off measurement in acute clinical practice. We have now rephrased and referenced this (paragraph 3 of Introduction)

Methods:
1. It is well-known that various states of alertness and agitation and fever affect the respiratory rate as well as its variability [Ref 10 of manuscript]. Clearly if children were recruited from all over the hospital, these simple parameters should have been recorded at least when the two study staff [RR2 and RR3] examine the patient. Thus it is clear that patients with bonds, head injuries, seizures, pain, diarrhea and vomiting for example, will have very variable for less than 30 seconds. Including this vast array of subjects is not helpful for the study and in fact completely influences the outcomes.

Thank you for these comments. When observer 2 and 3 took their measurements information was collected on the child's primary presenting complaint and a subjective assessment of their activity status was recorded. None of the other vital signs were collected as we felt these would not influence the comparison of RR measurements. Once the simultaneous measurement had been made, information was then collected on the RR1 measurement - method of measurement as well as the subjective assessment of the child's activity status by the HCP, asleep/awake/active.

In 26 of the measurements (15%) the subjective assessment of the child's activity during the measurement was different between the first and second/third RR measurements. From the analysis we were able to show that there was statistically no significant difference between these children whose activity had changed in between measurements. We have now included this further information within the results. This data can be found in paragraph 6 of the Results section. Below is a table breaking this information down and we could include this information within the paper if it is felt that this would add value.

Agreement of measurements based on child’s activity status			
Measurers	**Activity status**	**95% Limits of Agreement (Mean Difference)**	**Significance (p-value)**
RR 1 v RR 2	**Same activity status** (143 measurements)	-10.221 - 18.165 (3.972)	
	Discrepancy in activity status (26 measurements)	-9.658 - 14.899 (2.615)	p=0.269
RR 1 v RR 3	**Same activity status** (143 measurements)	-11.392 - 19.028 (3.812)	
	Discrepancy in activity status (26 measurements)	-11.329 - 17.252 (2.962)	p=0.210

All children recruited were stable on the wards and there were no acutely unwell children recruited. The clinical condition of the child had not changed in the period between measurements. This information has now been clarified in paragraph 2 of the Methods section. We do not feel that the clinical condition of the child would have had an effect on altering the RR between measurements and we have substantiated this by the fact that there was no significant difference observed in the pairwise agreements between measurements taken closer in time and those taken further apart. Below is a further table indicating this and again we could include this within the body of the paper if needed.

Table 4.8: Agreement and correlation of measurements by time taken

Measurers	**Time period**	**95% Limits of Agreement (Mean Difference)**	**Intraclass correlation coefficient (95% CI)**	**Significance (p-value)**
RR 1 v RR 2	**Early** - within 0-10 minutes (49 measurements)	-9.011 - 16.929 (3.959)	0.872 (0.681-0.939)	
	Late - within 20-30 minutes (69 measurements)	-9.623 - 15.652 (3.015)	0.899 (0.801-0.944)	p= 0.516
RR 1 v RR 3	**Early** - within 0-10 minutes (49 measurements)	-9.986 - 17.374 (3.694)	0.863 (0.697-0.931)	
	Late - within 20-30 minutes (69 measurements)	-9.790 - 17.123 (3.667)	0.869 (0.721-0.931)	p= 0.905

2. If these states have been collected, as they should have they should be presented at the analysis done taking this into account.

Please see comments above.

3. Table 1 should be corrected after careful perusal of reference 9.

This has now been changed to RR $\geq$ 95th centile as mentioned above.

Results:

1. While the age range is wide for children studied, there is no breakdown of the ages so that one can determine the usefulness.

We have now included a table of the age range of participants, please see Table 3 within Results section.

2. It is well-known that counting the respiratory rate for less than 60 seconds is quite inaccurate. It is not surprising that the nurses [R1] have the most variability, and with the least accuracy. Since most of the poor correlation refers to the R1 – R2 and R1 – R3 interobserver variability, this is not surprising at all and this does not had a call to the literature. All it tells us is that those hunting methods are poor.

Thank you for these comments. As mentioned above we feel the body of evidence previously available was very heterogeneous in its methods and reported findings. We have sought to provide substantial evidence now for these findings and to confirm the full extent of the inaccuracy in HCPs measurements. We wish to use our findings to attempt to highlight this issue further to all HCPs and thereby to improve practice either by improving clinical RR measurement education or by introducing new objective methods.

3. There is no description of how the results in table 4 were obtained

The results in Table 4 (now Table 5) were obtained after RR1 was taken. The HCP was asked the method which they used to take the measurement. This has been clarified in paragraph 5 of the Methods.

4. Assessment of tachypnea: for starters, the WHO algorithm should only be applied to children less than 59 months of age with cough or difficult breathing. Since most of these children do not have cough or difficult breathing [114 of the 169] and it is unclear which of the 55 with cough or difficult breathing are less than 59 months of age, the whole analysis of tachypnea means nothing.

We appreciate these comments and as such have changed this section of analysis to look at those children with a raised RR $\geq$ 95th centile, as discussed above. The Results and Discussion section now reflect this change.

Discussion:

1. Please see comment number one in the introduction relating to the first paragraph

Thank you - we have reworded the opening sentence of the discussion to reflect this.

2. on page 10, the authors refer to changes in the activity status of the child between measurements, not having an impact on the agreement. Perhaps this reflects that they did collect activity information. The data should be presented and analysis done by different activity states.

We have addressed this information above and included more data within the Results section.

3. Frankly all the R1 – R2 and R1 – R3 comparisons, only a reflection of the poor method for collection by the nurses and this should be emphasized rather than the comparisons. using a poor method of data collection does not justify the title "Poor interobserver agreement and the measurement of restoration children" in fact the interobserver agreement for the R2 – R3 comparisons was reasonable as those shown in several of the studies referred to above. In fact the agreement is quite good.

> *Thank you for this comment. We agree that the difference in agreement shown is mainly down to the poor methods used in clinical practice. We hope that this study has been able to emphasise this and what is happening at the frontline of clinical care. Indeed the agreement is much better for the simultaneous measurements which were completed under research conditions. We have shown that if HCPs were to use the recommended methods then the reliability of measurements can be improved greatly. We have reflected this statement in our conclusion.*

4. The whole discussion about tachypnea should be revisited. Any comparison with the WHO guidelines should be carefully re-examined.

> *We have since altered the discussion on tachypnoea as mentioned above.*

5. Most of the conclusions that the authors draw have been shown in several other studies over the years, but perhaps the most important observation that they have, is that if nurses are going to count the respiratory rate, they should be trained in doing it properly or not at all.

> *We agree with this and have re-emphasised our findings in the discussion and conclusion sections.*

VERSION 2 – REVIEW

REVIEWER	Burke, Derek Sheffield Children's NHS FT United Kingdom Competing interests: I work at the same trust as the authors and have undertaken some work on respiratory rate measurement with them. I have not been involved in this study.
REVIEW RETURNED	30-Jul-2017

GENERAL COMMENTS	An excellent paper which highlights the variability in measurement in respiratory rate using clinical methods even if the WHO recommended methodology is used. Of concern is that this variability is greater with abnormally high respiratory rates, calling into question the reliability of scoring systems using respiratory rates and the currently accepted "normal" values. I have no suggestions for improving the paper.

REVIEWER	Harron, Katie London School oy Hygiene & Tropical Medicine, UK Competing interests: no competing interests
REVIEW RETURNED	18-Aug-2017

GENERAL COMMENTS	This study attempts to evaluate the inter-observer agreement of respiratory rate count in children. The main flaw in this study is that

	the first “observer” is actually an unstated number of observers, comprising nurses and healthcare workers of different experience. Clearly there will be more variability between the measurements taken by multiple HCPs and each researcher, than between the two researchers. The implications of this need to be discussed. I have a few other points for clarification: Abstract: state the age range of children and the time period in which they were recruited. State how many different healthcare professionals took the first RR measurement. In the results, give the average RR counts taken by each group. Provide n/N for the 33% of children with agreement with RR>95th centile. Methods: The rationale for the r-z transformation is unclear. Please provide additional explanation. Explain how activity status was captured. Results: Absolute numbers /averages should be provided, e.g. where the different between measurements was assessed for different intervals of time, and by activity status. Provide row totals within tables. Table 4 – should be labelled as number of measurements taken by, not number of healthcare professionals (assuming there weren’t 169 different HCPs). The choice of denominator for the agreement for children with RR>=95% should be justified. It would seem to make more sense to base this on the number of children with RR>=95% according to the ‘gold-standard’ WHO criteria (observers 2 and 3) rather than any of the three observers, given the variability between raters. Discussion: Comment on how much a child’s RR is likely to change between the first and 2nd/3rd measurements.

VERSION 2 – AUTHOR RESPONSE

Reviewer 1 (Comments to the Author):

An excellent paper which highlights the variability in measurement in respiratory rate using clinical methods even if the WHO recommended methodology is used.

Of concern is that this variability is greater with abnormally high respiratory rates, calling into question the reliability of scoring systems using respiratory rates and the currently accepted "normal" values.

I have no suggestions for improving the paper.

We thank the reviewer for their generous comments and highlighting the importance of the findings from our work.

Reviewer 2 (Comments to the Author):

This study attempts to evaluate the inter-observer agreement of respiratory rate count in children. The main flaw in this study is that the first "observer" is actually an unstated number of observers, comprising nurses and healthcare workers of different experience. Clearly there will be more variability between the measurements taken by multiple HCPs and each researcher, than between the two researchers. The implications of this need to be discussed.

This has raised an important point. We opted for RR1 to be the HCP taking the RR as part of their normal clinical practice. The purpose of the study was to highlight the variability that currently exists within a clinical setting and therefore the intention was to include any HCP that may undertake respiratory rate measurements as part of their clinical practice and that the group as a whole would represent "current clinical practice". This would inevitably mean that we recruited a range of HCPs with different levels of experience. We have compared this variability with that of two researchers undertaking the readings under research conditions to highlight the variability that there is currently in clinical practice.

We feel that if we had used the same HCP for RR1 then this could have possibly altered their method of measurement and would not have truly reflected their actual practice. We did not want to introduce this potential bias by using the same HCP. We do agree that there could be more variability between the different HCPs taking RR1 by nature of their varied measurement techniques, and this is what we have emphasised within our findings.

As part of our analysis, which was not included in the final manuscript, we also analysed the difference in correlation and agreement in measurements taken by HCPs of different levels of seniority and experience. We found there that there was no statistically significant difference. Below is a table showing these results and we can include this in the body of the paper if the reviewer feels that this would provide greater clarity.

Agreement and correlation of measurements by level of seniority

Measurers	**Level of seniority**	**95% Limits of Agreement (Mean Difference)**	**Intraclass correlation coefficient (95% CI)**	**Significance (p value)**
RR 1 v RR 2	Band 5, HCW, Student nurse	-8.744 - 18.530 (4.893)	0.841 (0.570-0.923)	
	Band 6 and Band 7 nurse	-11.735 - 15.735 (2.000)	0.897 (0.827-0.938)	p= 0.150
RR 1 v RR 3	Band 5, HCW, Student nurse	-9.795 - 19.426 (4.816)	0.821 (0.576-0.908)	
	Band 6 and Band 7 nurse	-13.244 - 17.092 (1.924)	0.876 (0.796-0.925)	p= 0.219

Abstract:
1. State the age range of children and the time period in which they were recruited.

These alterations have now all been added to the Abstract section of the manuscript.

2. State how many different healthcare professionals took the first RR measurement.

Unfortunately, we were unable to keep a record of which healthcare professionals took RR1due to stipulations from the ethics committee who reviewed our study and so we do not have the data to quantify how many different HCPs were captured within RR1. We did record seniority levels however, as shown above.

3. In the results, give the average RR counts taken by each group. Provide n/N for the 33% of children with agreement with RR>95th centile.

Thank you for these suggestions, these alterations have now all been added to the Abstract section of the manuscript.

Methods:
1. The rationale for the r-z transformation is unclear. Please provide additional explanation.

The Fisher r-to-z transformation was used to assess any statistically significant difference between the correlation between different groups analysed, and allowed us to ascertain whether the difference between different selected groups was significant. On taking statistical advice we were advised that when correlation analysis is conducted on the same variables by two different groups, then the most appropriate way to do this is by transforming the correlation coefficients values into z scores.

2. Explain how activity status was captured.

When observer 2 and 3 took their measurements information was collected on the child's primary presenting complaint and a subjective assessment of their activity status was recorded. Once the simultaneous measurement had been made, information was then collected on the RR1 measurement - method of measurement as well as the subjective assessment of the child's activity status by the HCP, asleep/awake/active. Below is a table breaking this information down and we could include this information within the paper if it is felt that this would add value.

Agreement of measurements based on child's activity status

Measurers	**Activity status**	**95% Limits of Agreement (Mean Difference)**	**Significance (p-value)**
RR 1 v RR 2	**Same activity status** (143 measurements)	-10.221 - 18.165 (3.972)	
	Discrepancy in activity status (26 measurements)	-9.658 - 14.899 (2.615)	p=0.269
RR 1 v RR 3	**Same activity status** (143 measurements)	-11.392 - 19.028 (3.812)	
	Discrepancy in activity status (26 measurements)	-11.329 - 17.252 (2.962)	p=0.210

Results:

1.Absolute numbers/averages should be provided, e.g. where the different between measurements was assessed for different intervals of time, and by activity status.

Thank you for these suggestions, these absolute numbers have now all been added to the Results section of the manuscript.

2. Provide row totals within tables.

We have now included the total number of participants (n=) within each of the tables.

3. Table 4 – should be labelled as number of measurements taken by, not number of healthcare professionals (assuming there weren't 169 different HCPs).

This table has now been updated.

4. The choice of denominator for the agreement for children with RR>=95% should be justified. It would seem to make more sense to base this on the number of children with RR>=95% according to the 'gold-standard' WHO criteria (observers 2 and 3) rather than any of the three observers, given the variability between raters.

Thank you for this comment. In this section of the manuscript we were trying to analyse the agreement in assessing children with a higher RR. In doing so it was necessary to use measurements taken by both the WHO gold standard technique (used by observer 2 and 3) as well as different measurement techniques (often used by observer 1). This enabled us to ascertain whether the different measurement methods used were even more inaccurate in children with high RR.

Discussion:
1.Comment on how much a child's RR is likely to change between the first and 2nd/3rd measurements.

We have now provided extra information in the discussion section to address this point.

VERSION 3 – REVIEW

REVIEWER	Burke, Derek Sheffield Children's NHS FT Competing interests: I work at the same trust as the authors and have carried out work on respiratory rate monitoring with them previously.
REVIEW RETURNED	26-Sep-2017

GENERAL COMMENTS	An excellent paper which demonstrates significant inter-observer variation in the measurement of respiratory rate using manual methods. This variation is greatest at the higher respiratory rates raising concerns that significant physiological distress may not be licked up at an early stage. The authors advocate training for staff taking manual respiratory rate readings in clinical practice and also advocate the use of medical devices to record respiratory rate.

REVIEWER	Harron, Katie LSHTM, UK Competing interests: no competing interests

REVIEW RETURNED	26-Sep-2017

GENERAL COMMENTS	The authors have addressed my previous comments really well, but some of this information should be included in the manuscript, because other readers will likely have similar questions! Please include some discussion of the fact that you are comparing multiple reviewers (HCP) with single reviewers (RR1, RR2) and the implications for variability, and the comparisons between HCP-RR and between RRs only. It might be worth a sentence in the results stating that no differences were seen when stratifying by seniority. Average RR counts still need to be included in the results of the abstract, as does the denominator for children classified as raised RR. Thank you for explaining how activity status was captured. Again, just a sentence in the methods is needed, saying that this information was collected on the child's primary presenting complaint. Thank you for including the number of children in the stratified results sections; average RR values also need to be included.

VERSION 3 – AUTHOR RESPONSE

Reviewer 1 (Comments to the Author):

An excellent paper which demonstrates significant inter-observer variation in the measurement of respiratory rate using manual methods. This variation is greatest at the higher respiratory rates raising concerns that significant physiological distress may not be picked up at an early stage.

The authors advocate training for staff taking manual respiratory rate readings in clinical practice and also advocate the use of medical devices to record respiratory rate.

We once again thank the reviewer for their comments and for also highlighting the importance of the findings from our work.

Reviewer 2 (Comments to the Author):

The authors have addressed my previous comments really well, but some of this information should be included in the manuscript, because other readers will likely have similar questions!

We are pleased that we were able to address the reviewer's previous comments and hope that the additional changes to the manuscript enhance the overall message and findings of our paper.
Please include some discussion of the fact that you are comparing multiple reviewers (HCP) with single reviewers (RR1, RR2) and the implications for variability, and the comparisons between HCP-RR and between RRs only.

We have included in paragraph 4 of the discussion section an explanation that multiple HCPs were used for RR1 so as to reflect current clinical practice and included comments on the impact of this.

It might be worth a sentence in the results stating that no differences were seen when stratifying by seniority.

Thank you for this comment, we have included a sentence in paragraph 6 of the results section commenting on this.

Average RR counts still need to be included in the results of the abstract, as does the denominator for children classified as raised RR.

We have updated this section of the abstract and have included both the median RR counts for each of the observers and the denominator for the children who were classified as having a raised RR.

Thank you for explaining how activity status was captured. Again, just a sentence in the methods is needed, saying that this information was collected on the child's primary presenting complaint.

We have added in two further sentences within the methods section to clarify how the activity status was captured.

Thank you for including the number of children in the stratified results sections; average RR values also need to be included.

Thank you, we have now included the mean difference and 95% limits of agreement for those in the stratified results section, this can be found in paragraph 5 and 6 of the results section.