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) | Examining the criterion validity of two scalable, information technology-based systems designed to measure the quantity and quality of movement behaviors of children from Hong Kong primary schools: A cross-sectional validation study |
|---------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AUTHORS | Ha, Amy S.; Cheng, James; Chan, Cecilia; Jiang, Guanxian; Yang, Yijian; Ng, Johan Y Y |

**VERSION 1 – REVIEW**

| REVIEWER | Xiong, Xiuqin |
|----------|--------------|
| REVIEW RETURNED | 12-Feb-2022 |

| GENERAL COMMENTS | General comments: This study provided validity evidence for two systems designed to measure children's moderate-to-vigorous physical activity and fundamental movement skill proficiencies. It is generally an interesting paper. I have some confusions in some places. Please see some comments and questions as below. Major comments: Page 4 line 10 1. The authors said they used well-established tools to generate criterion measure for validation purposes. I am not very experienced in this area. In other areas, the validation process may include for example, known-group validity, convergence validity and test-retest reliability, etc. Could you please state what are the well-established validation method in your area and do you cover them all or only part of them? I can see your paper include correlation and kappa. It would be great if this can be added in the method part or as an appendix. Method part: Page 10, line 24-40 2. There seems no description of the scores of accelerometer and Sport Bands used for the Pearson correlation. Are they continuous? What are the range of these scores? Are higher scores indicating better or more PA? 3. If using Pearson correlation, it would be great to state the reason for choosing this method. There are some assumptions of using Pearson correlation, such as being normally distributed. If not, maybe change to Spearman correlation. 4. For the kappa calculation, could you please describe what the variable looks like? Is it binary or multiple-level variable? If it is multi-level categorial variable, weighted kappa can also be considered. In addition, please add reference for the interpretation of kappa values. (page 10, line 39-40) Results: page 12, line 40-58 |
5. “The additional analysis: association between sport band and …” seems not have been mentioned in the methods data analysis part and I wonder why this analysis is needed. Could you add this in the method? (although I see some explanation in the discussion)
6. I did not quite understand how correlation between sport band and FMS demonstrate the construct validity. Could you please explain a bit? In addition, the correlation seems quite low (0.217), right?
7. Could you please explain more about how does the fundamental movement skill scores were standardized by skill, grade and sex? (if not enough space in main paper, appendix would also be great)
Discussion: page 14, line 36-40
8. The sample size was only 25, which was mentioned as a limitation. I understand that research during Covid can be difficult, but could you please explain or provide evidence if this (25) meets the minimum sample size for statistical analysis in your study and how does the small sample size can influence your result and interpretation?

Minor comments:
Would it be possible to split the introduction into two or three instead of one big long paragraph? This may make it clearer.

REVIEWER
Reed, Julian A.
Furman University
REVIEW RETURNED 16-Mar-2022

GENERAL COMMENTS
Overall this is a good paper and can contribute to the knowledge base in this area. Recognizing links between motor ability and PA is important. Easily wearable devices will continue to progress and get more ubiquitous. The need for simple, relatively inexpensive devices will only help to collect valid and reliable data.

I think the authors need to provide much more information on how Raters were trained. "experts" is not enough. A detailed process should be provided and if there were any raters who do meet the minimum inter-rater reliability to be included.

The sample size was small and limits external validity, however if revised to demonstrate this was a pilot study would be acceptable.

More information needs to be provided on the standardization process of the FMS’s. What was the hypothesis for the additional analyses between Sports Band and the Rater Scores? Are the authors suggestion no associations between MVPA and FMS? Does this need to be included since it was not purpose of the paper.

The paper is well written, but needs further details on the aforementioned comments for acceptance.

VERSION 1 – AUTHOR RESPONSE
Reviewer: 1
General comments:
This study provided validity evidence for two systems designed to measure children’s moderate-to-vigorous physical activity and fundamental movement skill proficiencies. It is generally an interesting paper. I have some confusions in some places. Please see some comments and questions as below.

Authors’ response:
We would like to thank the reviewer for showing interest in our work. We also hope that our revisions and responses could eliminate some confusions the reviewer might have had when reading the previous version of our manuscript.

Major comments:
Page 4 line 10
1. The authors said they used well-established tools to generate criterion measure for validation purposes. I am not very experienced in this area. In other areas, the validation process may include for example, known-group validity, convergence validity and test-retest reliability, etc.

Authors’ response:
The main type of validity we examined in this study was criterion validity. We acknowledge that there are slight deviations in terms of how various aspects of validity are defined in different fields, but we feel the term “criterion validity” would be accepted by most researchers across different fields. We have made clarifications for this in the title, article summary, and throughout the manuscript.

Could you please state what are the well-established validation method in your area and do you cover them all or only part of them?

Authors’ response:
We have stated the tools that were used as the criterion measure in our study in this section.

“Used well-established tools (i.e., accelerometry for physical activity, expert ratings of recorded fundamental movement skills) to generate comparative measure to examine criterion validity of developed tools” (page 3 lines 48-50)

I can see your paper include correlation and kappa. It would be great if this can be added in the method part or as an appendix.

Authors’ response:
We have provided the details of our data analytical approaches under the “Data analyses” sub-section within the “Method” section in the previous version of the manuscript. We have also explained how correlations and kappa coefficients were calculated and used in the study.

“Pearson correlation was calculated to determine the strength of the linear association between MVPA scores captured using accelerometers and Fun to Move@JC Sport Bands.” (page 10 lines 226-227)

“In terms of FMS data, the percentage agreement and kappa coefficients for criteria of each skill (parameterized as 0 = criterion not met; 1 = criterion met), scored by the FMS and expert raters respectively, were calculated. The kappa coefficient accounts for expected agreement by chance and hence is a more robust measure for examining agreements” (page 10 lines 229-231)

Method part: Page 10, line 24-40
2. There seems no description of the scores of accelerometer and Sport Bands used for the Pearson correlation. Are they continuous? What are the range of these scores? Are higher scores indicating better or more PA?

Authors’ response:
In response to the reviewer’s feedback, we have revised the section to clearly define the unit of analyses, and how values of these outcome should be interpreted.

“The unit of output and analyses was children’s time spent in MVPA per calendar day. Therefore, larger values for this continuous variable represent engagement in more MVPA. Data collected using both sets of devices over a 7-day period was used for comparison purposes.” (page 9 lines 193-196)

We have also additional descriptive statistics on these scores following feedback from the reviewer:

“… the mean MVPA measured using accelerometers and Sport Band were 43.8 ± 25.0 mins and 49.2 ± 20.4 mins, respectively.” (page 12 lines 259)

3. If using Pearson correlation, it would be great to state the reason for choosing this method. There are some assumptions of using Pearson correlation, such as being normally distributed. If not, maybe change to Spearman correlation.

Authors’ response:
Pearson correlation was chosen in this study because we were interested in examining the linear relation between the two measures. This justification was provided in the revised manuscript:

“Pearson correlation was calculated to determine the strength of the linear association between MVPA scores captured using accelerometers and Fun to Move@JC Sport Bands.” (page 10 lines 226-227)

Nonetheless, non-normality was indeed observed in our data, and therefore we have also added information regarding the Spearman correlation in the revised draft:

“The Pearson correlation between accelerometer- and Sport Band-measured MVPA was \( r = .795 \) (\( p < .001 \)). However, results of the Shapiro-Wilk test suggested both sets of data were non-normal (\( ps < .05 \)). As such, the Spearman correlation between the two measures was calculated. A significant correlation between was also found with Spearman’s \( \rho = .460 \) (\( p = .021 \)).” (page 12 lines 261-265)

We have decided against removing the Pearson correlations because our primary focus is the linear relation between the variables, while Pearson correlations is considered by many researchers as a metric robust to non-normality. As such, we have presented the Spearman correlation as a supplementary result.

4. For the kappa calculation, could you please describe what the variable looks like? Is it binary or multiple-level variable? If it is multi-level categorial variable, weighted kappa can also be considered. In addition, please add reference for the interpretation of kappa values. (page 10, line 39-40)

Authors’ response:
The variable is indeed a binary variable. In response to the comment made by the reviewer, we have now added information regarding how the variable was parameterized in the manuscript. We have also inserted the reference regarding how kappa values should be interpreted.
“... the percentage agreement and kappa coefficients for criteria of each skill (parameterized as 0 = criterion not met; 1 = criterion met), scored by the FMS and expert raters respectively, were calculated.” (page 10 lines 230-231)

Results; page 12, line 40-58
5. “The additional analysis: association between sport band and ...” seems not have been mentioned in the methods data analysis part and I wonder why this analysis is needed. Could you add this in the method? (although I see some explanation in the discussion)

Authors’ response:
We have re-ordered the text and supplemented the data analysis section with some additional information:

“As an exploratory attempt, we also examined the relation between children’s Sport Band-measured MVPA and FMS scores derived from the developed rating system. Data for this analysis was drawn from the cloud storage where data collected both systems were synchronized to... Pearson correlation between MVPA and FMS scores was then computed to examine the linear relation between these two outcomes.” (page 11 lines 235-246)

6. I did not quite understand how correlation between sport band and FMS demonstrate the construct validity. Could you please explain a bit? In addition, the correlation seems quite low (0.217), right?

Authors’ response:
In our introduction section, we mentioned that some researchers have shown that FMS proficiency is related to MVPA outcomes in children. With the assumption that these constructs are indeed associated, a significant correlation between Sport Band and FMS Rater-measured outcomes would be evidence of construct validity. However, since (a) the relation between these constructs is not extremely clear cut in the extant literature, (b) the correlation found was not significant / strong, and (3) considering the comments made by the reviewer, we have toned down this statement and have removed our argument regarding construct validity. Instead, we have presented this comparison as analyses conducted exploratorily.

7. Could you please explain more about how does the fundamental movement skill scores were standardized by skill, grade and sex? (if not enough space in main paper, appendix would also be great)

Authors’ response:
The following description has been added to the manuscript:

“To account for differences in terms of children’s attributes and skill difficulty, FMS scores were standardized by skill, and by children’s grade and sex prior to analyses. Specifically, the score of each FMS assessment was converted to a z-score based on the child’s grade level, sex, and the skill. An aggregated score for FMS was then calculated by taking the mean of z-scores from all assessments.” (page 11 lines 240-244)

Discussion: page 14, line 36-40
8. The sample size was only 25, which was mentioned as a limitation. I understand that research during Covid can be difficult, but could you please explain or provide evidence if this (25) meets the minimum sample size for statistical analysis in your study and how does the small sample size can influence your result and interpretation?

Authors’ response:
First, we wish to note that the correct sample size should be 75, but not 25. We sincerely apologize for the mistake and have corrected this in the manuscript.

Secondly, the direct consequence of the relatively small sample size is that our analyses may lack statistical power. Therefore, the results should be interpreted with caution. As such, the following sentences have been inserted to the corresponding section of the manuscript.

“Results from a post-hoc power analysis suggested that the observed power for this analysis was .455, falling short to the common benchmark of .80. As such, the results of the current study should be interpreted with caution.” (page 15 lines 349-351)

Minor comments:
Would it be possible to split the introduction into two or three instead of one big long paragraph? This may make it clearer.

Authors’ response:
We would like to thank the reviewer for this suggestion. We have now broken the opening paragraph into two.

Reviewer: 2
Comments to the Author:
Overall this is a good paper and can contribute to the knowledge base in this area. Recognizing links between motor ability and PA is important. Easily wearable devices will continue to progress and get more ubiquitous. The need for simple, relatively inexpensive devices will only help to collect valid and reliable data.

Authors’ response:
We would like to thank the reviewer for acknowledging the importance of our work, and the potential implications to future research.

I think the authors need to provide much more information on how Raters were trained. "experts" is not enough. A detailed process should be provided and if there were any raters who do meet the minimum inter-rater reliability to be included.

Authors’ response:
First, we would like to clarify that both expert raters were experienced in assessing FMS prior to this study. Nonetheless, in response to the comments made by the reviewer, we have included some additional information regarding the training and reliability checking procedures that were employed initially:

“Specifically, both raters received a one-hour training by an author of the paper. During the training session, they were introduced to general methods of conducting process-based assessment of motor skills and were guided through the assessment criteria for all skills. At the end of the training, the raters were provided with, for each skill, eight to ten video clips of children’s FMS performance. The trained raters were then asked to rate all skill performances based on the set criteria individually. Their scores were then compared with a set of results that were agreed between multiple authors of the study. Accordance with previous practice, both raters have established over 90% coding reliability with the authors when applying the TGMD-3 … Since some assessment criteria were modified to align with those used in the FMS Rater in this study, a pilot test was conducted specifically for this study. The two raters each rated 20 children performances per skill independently, and their scores reached an agreement above 90% for each skill.” (page 10 lines 208-221)
The sample size was small and limits external validity, however if revised to demonstrate this was a pilot study would be acceptable.

Authors’ response:
We acknowledge that the sample sizes for some of our analyses may appear to be relatively small, and consequently may limit external validity. We have added this point in our discussion section:

“Second, the sample sizes for some analyses in our study were relatively small. Despite observing strong correlations between Sport Band-measured MVPA and the criterion measure, there is a need to expand the pool of participants in terms of age and PA levels. Repeating the procedures in a wider of participants will also expand the external validity of our findings.” (page 15 lines 343-347)

However, apart from the examination between MVPA and FMS scores (which we now positioned it as exploratory in nature), our results suggest that the analyses are sufficiently powered. As such, we are slightly hesitant in naming our entire study (involving over 1,000 children) a pilot. Nevertheless, if the editor and reviewer both agree this is the best way to move forward with our manuscript, we will comply to this suggestion.

More information needs to be provided on the standardization process of the FMS's.

Authors’ response:
In response to the reviewer’s comment, we have added the following description to our manuscript:

“To account for differences in terms of children’s attributes and skill difficulty, FMS scores were standardized by skill, and by children’s grade and sex prior to analyses. Specifically, the score of each FMS assessment was converted to a z-score based on the child’s grade level, sex, and the skill. An aggregated score for FMS was then calculated by taking the mean of z-scores from all assessments.” (page 11 lines 240-244)

What was the hypothesis for the additional analyses between Sports Band and the Rater Scores? Are the authors suggestion no associations between MVPA and FMS? Does this need to be included since it was not purpose of the paper.

Authors’ response:
Given the mixed results found in past studies, we felt there was probably insufficient evidence for us to formulate a strong hypothesis. As such, we have now modified our paper to state that this was an exploratory attempt to look at the potential association between MVPA and FMS measured using the developed systems.

In terms of the result of the analysis, we believe our findings (i.e., \( r = .22, p = .067 \)) could be interpreted multiple ways (e.g., no association, marginally significant), and perhaps to some degree depending on the stance of the subject. We feel the most honest and informative way to present the results is to provide the statistics as they are.

Since the relation between these constructs is of interest to many researchers, we feel there is a need to present these findings in the paper, despite it not being the main purpose of the paper. Although we no longer claim that these results can be used as evidence to support the validity of the systems, we feel this could potentially instill confidence in other researchers that this is a potential way to further examine the relation between the constructs in future research.

The paper is well written, but needs further details on the afomentioned comments for acceptance.
Authors’ response:
Once again, we would like to thank the reviewer for the positive feedback. We hope the responses we provided above have sufficiently addressed the concerns of the reviewer.

VERSION 2 – REVIEW

| REVIEWER       | Xiong, Xiuqin                  |
|----------------|--------------------------------|
|                | The University of Melbourne    |
| REVIEW RETURNED| 27-May-2022                    |
| GENERAL COMMENTS| Thank you for the response and revision of the manuscript according to my previous comments. No further comments. |