-> We thank the reviewer 1 for constructive comments and suggestions on the manuscript and addressed them all in the revised version:

Anonymous Referee #1
General Comments This is an interesting paper that examines the ability of farmers/students/experts in carrying out a qualitative test for soil moisture. Although the sample is small, the results are encouraging particularly when training is provided. The potential of this approach for soil moisture assessments is clearly acknowledged at the end of the paper, i.e. upscaling and data transmission via SMS. Overall the paper makes a good contribution to the literature.

Specific Comments
1. Some reference to the literature on citizen science and more
recent attempts at involving students in collecting soil related information should be added, e.g. the OPAL initiative, GLOBE, etc. 2.

-> We included a paragraph about citizen science and crowed-sourcing initiatives (e.g. OPAL and GLOBE) in the conclusions showing the potential application of the qualitative scheme

2. What is the source of soil information in the pilot area? Existing soil maps? A survey undertaken by the authors?

-> The soil information stated in the discussion paper was determined in the course of an analysis in the field. In the meantime additional lab analysis and discussion with soil scientists resulted in the following soil classification:
Profile 1: Chromic Cambisol Colluvic Clayic / Profile 2: Haplic Cambisol Siltic Ruptic. (WRB, 2014). The texture is ranging from clay to loamy sand. Further soil information is updated in the site description section (chapter 2.2) and citation is given.

3. What is the soil classification system used, e.g. WRB 2006, WRB 2014 as the combination of Haplic Andosol, loamic, fluvic does not conform to any combination of Reference Soil Group or Qualifiers in these systems. Please explain.

-> We used the WRB (2014) and stated this in the text (site description, chapter 2.2) and list of reference. The soil description in the first manuscript was preliminary, but after final evaluation of all data was updated to Chromic Cambisol Colluvic Clayic, according to WRB (2014).

4. Soil moisture and volumetric content of water in the soil are closely related to soil texture. What is the range of soil textures in the plots?

-> The texture is in the range of clay and clay loam, with intercalated layers of loamy sand. Texture classes are given in the site description and we added further information on how the texture influences hydraulic properties like wilting point and field capacity.

5. Please clarify how you mapped the measured volumetric content of water to your soil moisture classes, e.g. did you use the median of the estimate soil moisture classes to do the assignment?

-> For Fig. 3a and 3b: Corresponding qualitative wetness classification were made by the first author at the same time the gravimetric samples were taken to avoid the influence of a potential drying effect as sampling was slow and took longer than the qualitative test with the farmers, students and experts.
We added this sentence at the end of section 2.2.

6. More detail should be added to the description of your wetness classification scheme. Although you refer to a previous paper, you also refer to a modification and this should be explained here in more detail.

-> The modification was necessary to account for local peoples' every-day experience which, in Tanzania, is more related to farming and brick making but not to hiking and outdoor recreation activities like in Switzerland. We stated this clearly at the end of the introduction section (chapter 1).
We reformulated and extended the section in chapter 2.1 on the description of the wetness classes.

7. The soil moisture of the uppermost layer is not representative of the whole soil profile. How do you know these samples were at equilibrium? Think about replacing the outdated reference of 1927 to something more recent.

-> The thought was, that it might be possible to anticipate wetness at root depth knowing the characteristics of the water retention curve but we skipped that part in the method and discussion section and recommended to use the qualitative classification scheme in combination with the "Spade Method" (Görbing, J. & Sekera, F., 1947).
We added a more recent citation for crop root depth but decided to keep the 1927 reference because a lot of work on crop physiology has been done early in the 20th century and is still valid. To clarify this part in the revised version of the manuscript, we rewrote and added a few related statements.

8. Add a reference to the Mann-Whitney test and Bonferroni significance.

-> We added the original reference

9. Although the test appears to be visual, it also involves removing some of the top soil. If this is being done multiple times by the farmers/experts/students, does this not affect the result due to disturbance?

-> Yes, if every farmer would have removed the uppermost soil that would have be misleading. People were instructed accordingly and only the first participant actually removed the uppermost soil. All following participants were assessing the soil sample at the soil surface.
To avoid confusions, we skipped the part at the end of section 2.1.

-> We thank the reviewer 2 for constructive comments and suggestions on the manuscript and incorporated the feedback in the new version of the text:

Anonymous Referee #2
General Comments:

1) This manuscript describes the testing of a soil “wetness” classification scheme with three groups of individuals (experts, students, and farmers) in Tanzania. The methods and results presented here have importance for both “experts” and farmers that seek to determine optimal conditions for seeding and for maintaining crop vigor. The paper was well-written and straight-forward. One point in particular that I feel warrants further discussion is how transferable the wetness scheme and training is between sites/soils.

-> We highlighted the importance of calibration to the local soil type and incorporating the cultural setting in the conclusions.

2) For example, the classification scheme refers to conditions in which bricks could be made, which appears to be a highly localized bit of knowledge. Overall, it seems that this process involved a fair degree of “calibration” to local practices and soil conditions, so how dependent is the process on the experts coming into a given locale?

-> The intention of the method is to incorporate the tacit knowledge of local people and from the interviews with local farmers the criterion “brick making’ and “seeding” turned out to be the common and widespread activities with what local people are familiar with in terms of soil moisture. While knowing that our interviews were restricted to farmers of this test sites near Arusha, we would assume, that this is similar in a lot of rural communities in semi-arid regions.
After extensive field testing it may be possible to develop a genre of calibration methodology that can easily be used as a step by step guide for local agricultural extension officers for example.
Concerning the dependency of the actual volumetric soil water content and the qualitative wetness classes we fully agree, that this is subject to calibration to the local soil. In fact we mention in the paper, that the volumetric water content of the driest wetness class in the Swiss study was similar to the volumetric water content of the wettest class in the African study. In many applications however we think, it is more important to assess whether one spot is drier or wetter than a different one (relative difference). Rinderer et al. (2012) and this paper could at least show that this is possible for two contrasting environments.

3) Similarly, could these methods be developed into fact sheets or training guides that could be used by non-expert individuals in multiple locations (in theory at least)?

-> We see potential to use the method for soil moisture assessment in different locations why we provided the assessment forms as templates in the supplement material.

4) I wonder why there was no testing (or at least discussion) of whether these practices resulted in any kind of improvement in farming practices and/or yields. That seems to be the ultimate goal of the effort, yet as far as I can tell was not mentioned.

-> The project was only a pilot study with limited duration and therefore it was not possible to systematically analyses impacts (benefits) on crop yields or farming practices. The authors however have handed in a project proposal, with one of the work packages is involved with quantifying the value of soft soil moisture information on village communities, individual farmers and decision makers. This proposal is currently subject to a review process.
We added a sentence at the end of the conclusions.

5) I am a bit confused by what constitutes the “control” in these experiments. Is there a single person who decides the “true” wetness classification of the sampling points, or is that determination reliant on the median response from the group?

-> For the assessment of the robustness of the method i.e. if different individuals come up with a very similar or very different classification, the agreement among the individuals was assessed and we chose the median classification as “control”.
For the assessment of the agreement between qualitative and quantitative moisture content we used gravimetric samples which took longer than the test with the individuals. To avoid the influence of a potential drying effect during this period we chose the classification of the first author, who did the rating at the time of gravimetric sampling.
We mentioned this explicitly in the method section.

6) Based on the limited differences between volumetric water contents in the dry (1 & 2) and wet (5-7) classes, this seems like important information when judging the accuracy of responses.

-> We extended the text on the pros and cons for merging classes in the discussion section. We added a sentence that qualitative differences are not necessarily reflected in quantitative differences.

7) Finally, the authors argue against reducing the number of wetness categories. However, Figure 3 in particular leads me to question the value of having so many different classes, when there is little or no quantitative moisture differences between many of them.

-> While we agree in cases where the volumetric water content is relevant, we also see applications where the qualitative information is of importance (e.g. mapping saturated areas with overland flow versus wet areas with shallow subsurface flow; (see Brazkova et al (2012) or Ali et al (2013) or Dunne and Black 1960).
We extended the text to be clearer about this in the discussion section.

8) At the same time, even after training many of the farmers appear to have misclassified samples by multiple categories (Figure 8), which seems to suggest that there may be little benefit in having seven different classes.

-> Figure 8 shows the mean classification difference for all sampling points of each wetness class per test person. In other words: Is an individual person systematically classifying all sampling points of a given wetness class too dry or too wert? During the first test (figure 8a) we agree with the reviewer, that misclassification by several classes was given but figure 8b

shows that after a longer introduction none of the wetness classes is systematically classified too wet or too dry by more than one or two classes (see pastel colors).
We added an example in the figure captions to be clearer to the reader.

Specific Comments:

9) P3035, L3: How realistic is the assumption that vertical soil moisture is close to equilibrium? This assumption seems suspect to me, but could possibly be verified by repeated tests through time or depth. A citation here would help.

-> While we agree that entire soil profiles under natural conditions are often not in equilibrium, we think that this held within the uppermost 15 cm where we classified the soil wetness and took the quantitative measurements. This is the most relevant layer for seeds and young crops with a shallow root depth. To avoid misunderstanding we however removed the sentence in section 2.1 and instead recommend to use the “Spade Method” if soil moister at depth is of importance. (Görbing, J. & Sekera, F., 1947). To clarify this part in the revised version of the manuscript we rewrote and add a few related statements.

10) P3038, L5: It is not clear why maximum attainable CK value (CKmax) would be less than 1, since the previous statement states that perfect agreement would be CK = 1. What is the value of CK/CKmax if CKmax is not a constant?

-> CKmax is smaller than 1 in cases where the codes are not equally probable and not both raters assign all classes similarly often (Sim and Wright, 2005). So CK/CKmax expresses the obtained CK relative to the maximum obtainable CKmax (given that the marginal distributions are not equal) which allows to compare this measure among individuals and among the individual tests.
We rewrote this paragraph to be clearer.

Sim, J.; Wright, C. C. (2005). "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements". Physical Therapy 85 (3): 257–268.

11) P3041, L3-13: How were the samples divided into “halves”? The meaning of this entire paragraph was not clear to me.

-> We analyzed the inter-rater reliability for the first half of the points along the parcours (site 1-20) and the second half of the parcours (site 21-40).
We clarified this in the text (last paragraph of section 3.2).

12) P3042, L8: Practically speaking, is there any difference in outcome if a farmer rates a soil as too wet versus too dry?

-> One of the potential applications of the scheme is, that farmers can compare soil wetness among their fields and have a measure to tell whether the own or the neighbor’s field needs water first. Farmers in the case study sites already use some sort of qualitative soil moisture assessments to decide when a field needs irrigation water or not but these are quite subjective and a more objective method would help to prevent individual biases. At least the acceptability of this method in a village will depend on the robustness of the scheme.

13) P3043, L5: I wonder why this information about the misclassification of 6 classes due to ticking error was not included in the results section, since on P3041, L1 there is mention of confusing assessment form for the April test.

-> In the results we talk about classification off by more than four classes but we included the numbers of misclassifications of 5 and 6 classes explicitly.

14) P3044, L6-8: Since the median volumetric water contents were practically identical between the two driest classes and three wettest classes, how does one resolve small scale changes in soil moisture using them? Or are the classification schemes capturing differences in the soil matric potential, even if they do not appear to be significant changes in soil moisture?

-> We assume, that the simple method is not capable to capture changes in matric potential. However we think, that the small scale pattern of e.g. saturated versus not saturated (but very wet sites) can be a useful information in many applications.
We changed the text to be clearer.

15) P3045, L7-8: In my opinion, the line “The study also shows that the qualitative wetness classes are reflecting the quantitative differences in volumetric water content” is debatable, since the results and discussion both reflect that many of the wetness classes did not have statistically significant differences between them. For example, Figure 3 indicates that Class 6 may have had slightly higher VWC values compared to Class 7.

-> We are more precise in stating, that we think it is only valid for the median values of the intermediate wetness classes but not for the dry and wet classes.

16) Table 1: The statement “water liquefies” does not make sense to me. Water is generally liquefied.

-> While we see some potential to better phrase this, we have to give the original class description, which was given on the forms used during the test.

17) Figure 4: It seems like this could be put into a single chart, since there appears to be mostly repeated data between a) and b). At the same time, it seems like this information is wholly contained within Figure 6 (though certainly in a different form). Is the figure necessary?

-> We combined the two figures and included the difference between students with basic introduction and with training.

18) Figures 6, 7, and 8: It took me a while to interpret these figures. Maybe the caption could be improved to better convey how to interpret the information (possibly by using an example sampling point and an example test person).

-> We included an example in the figure captions of figure 6, 7, 8