Agron. Sustain. Dev. (2017) 37:32
DOI 10.1007/s13593-017-0441-y

RESEARCH ARTICLE

The accuracy of farmer-generated data in an agricultural citizen science methodology

Jonathan Steinke[1,2] · **Jacob van Etten**[1] · **Pablo Mejía Zelan**[3]

Accepted: 30 June 2017

Abstract Over the last decades, participatory approaches involving on-farm experimentation have become more prevalent in agricultural research. Nevertheless, these approaches remain difficult to scale because they usually require close attention from well-trained professionals. Novel large-*N* participatory trials, building on recent advances in citizen science and crowdsourcing methodologies, involve large numbers of participants and little researcher supervision. Reduced supervision may affect data quality, but the "Wisdom of Crowds" principle implies that many independent observations from a diverse group of people often lead to highly accurate results when taken together. In this study, we test whether farmer-generated data in agricultural citizen science are good enough to generate valid statements about the research topic. We experimentally assess the accuracy of farmer observations in trials of crowdsourced crop variety selection that use triadic comparisons of technologies (tricot). At five sites in Honduras, 35 farmers (women and men) participated in tricot experiments. They ranked three varieties of common bean (*Phaseolus vulgaris* L.) for *Plant vigor*, *Plant architecture*, *Pest resistance*, and *Disease resistance*. Furthermore, with a simulation approach using the empirical data, we did an order-of-magnitude estimation of the sample size of participants needed to produce relevant results. Reliability of farmers' experimental observations was generally low (Kendall's *W* 0.174 to 0.676). But aggregated observations contained information and had sufficient validity (Kendall's tau coefficient 0.33 to 0.76) to identify the correct ranking orders of varieties by fitting Mallows-Bradley-Terry models to the data. Our sample size simulation shows that low reliability can be compensated by engaging higher numbers of observers to generate statistically meaningful results, demonstrating the usefulness of the Wisdom of Crowds principle in agricultural research. In this first study on data quality from a farmer citizen science methodology, we show that realistic numbers of less than 200 participants can produce meaningful results for agricultural research by tricot-style trials.

Keywords Citizen science · Crowdsourcing · Wisdom of Crowds · Participatory methods · Participatory variety selection · Common bean

1 Introduction

Agricultural research has increasingly incorporated participatory methods over the last decades in order to become more client-oriented, addressing the variable conditions and preferences of resource-poor farmers and consumers (Lilja et al. 2013). Farmer participation in agricultural research has not yet become mainstream throughout the world, however. Scaling of participation is hindered not only by institutional constraints and prejudices about data quality but also by the resource-intensity of most participatory methodologies in terms of time, training, and cost per beneficiary (Hellin et al. 2008; Morris and Bellon 2004). The scalability of current participatory approaches is often limited because they rely

Electronic supplementary material The online version of this article (doi:10.1007/s13593-017-0441-y) contains supplementary material, which is available to authorized users.

✉ Jonathan Steinke
j.steinke@cgiar.org

[1] Bioversity International, c/o CATIE, Turrialba 7170, Costa Rica

[2] Humboldt-Universität zu Berlin, Division of Horticultural Economics, Unter den Linden 6, 10099 Berlin, Germany

[3] Programa de Reconstrucción Rural, Horconcito, Santa Barbara, Honduras

Published online: 24 July 2017

on organized farmer groups that require high levels of professional support and the establishment of collective plots (cf. Witcombe et al. 1996 for participatory variety selection). Moreover, these approaches require that farmers stay engaged throughout the crop cycle, even when participation involves hard work, for example, weeding the plot. This often leads to a free-rider problem. Limited participation due to free-riding has led to incomplete observations, haphazard crop variety choices, and subsequent dis-adoption of selected varieties, when farmers discover the true field performance of the variety on their own plot of land (Misiko 2013).

Alternative decentralized approaches to farmer-participatory research have been suggested recently, emphasizing linkages between farmers within research, knowledge, and innovation networks (Desclaux et al. 2008; Spielman et al. 2009). Such systems can lead to success (e.g., Humphries et al. 2015), but farmer networks require long-term commitment and cost-intensive facilitation by outsiders, and results are difficult to scale beyond the local level (Classen et al. 2008). Because increasing both empowerment and the numbers of farmers as beneficiaries are goals in participatory research (Snapp and Heong 2003), there is still an open need for methodologies that include farmer-led innovation processes *and* that are scalable.

Citizen science and "crowdsourcing" methods can help to overcome the limited scalability and free-rider problem in existing participatory methodologies for the agricultural sciences. Large-*N* citizen science projects can involve large groups of volunteers individually contributing to scientific tasks, notably data collection, by crowdsourcing approaches (Dickinson et al. 2012; Hand 2010). In these approaches, large research tasks are first subdivided into many "micro-tasks" that are doable for an individual participant. These tasks are distributed and results are collected through digital channels. The individual results are then combined to produce a large result. Crowdsourcing has been used for many applications, from translation to image recognition. When crowdsourcing involves the production of new knowledge, it relies on the "Wisdom of Crowds" principle (Surowiecki 2005). This principle implies that large groups of participants can in aggregate produce highly accurate results when certain conditions are met: a sufficient diversity of viewpoints and independence of observations. Use of information and communication technologies to receive contributions from many participants makes citizen science research scalable (Dickinson et al. 2012). As research relies on individual rather than group contributions, it also avoids the free-rider problem.

Here, we study the large-*N* citizen science approach proposed by Van Etten (2011) in an application to the detection of phenotypic differences between varieties of common bean, *Phaseolus vulgaris* L. (Sect. 2.1). While the full methodology involves more steps, such as participatory research priority setting, we here focus on farmer observation as a method of data collection. Large-*N* citizen science can yield accurate results if the relatively low reliability of farmers' individual observations is balanced by a large number of observations, following the Wisdom of Crowds principle. To achieve high-quality science using this citizen science approach, the level of accuracy of data collection by farmers should be clear. Farmers have heterogeneous levels of knowledge and expertise, and possess knowledge that is more developed in some domains than in others (Bentley 1989). Kremen et al. (2011) show that citizen scientists can make accurate observations in certain categories, but are more prone to bias or inaccuracy in others. Therefore, the first goal of our study is to assess the accuracy of farmers' observations in citizen science trials, as a proof of concept. Feasibility also depends on the number of participants that are needed to achieve accurate results. Therefore, the second goal of our study is to gain insights regarding the order of magnitude of the number of participants required to produce useful findings with a large-*N* citizen science methodology.

2 Materials and methods

2.1 Citizen science methodology

We apply a citizen science approach first proposed by Van Etten (2011). The approach is based on triadic comparisons of technologies, which we refer to here as the *tricot* approach. Tricot can be used to assess a range of agricultural technologies. When it is applied to crop varietal tests, women and men farmers each receive experimental seed quantities of three different varieties chosen randomly from a larger set of varieties, and grow these varieties next to their own crop, under usual crop management (Fig. 1). Farmers observe the three varieties and evaluate different aspects of their performance at different points in time, using a simple ranking format, triadic comparisons. Farmers then communicate their observations to field agents verbally, on paper, or via mobile telephone. The farmer-generated observation data are analyzed using statistical methods for ranking data. Given an adequate number of partial rankings, a preference scale for all varieties included in the experiment can be constructed by fitting a Bradley-Terry (BT) model (Bradley and Terry 1952; Coe 2002). Also, more sophisticated models for preference data can be used (e.g., Fürnkranz and Hüllermeier 2010; Strobl et al. 2011). Early experiences with applications of tricot are described by Van Etten et al. (2016).

Triadic comparisons are a proven method in ethnobiological research (Martin 2004). This format allows farmers to register and communicate their observations with a low level of literacy, and without the need to make quantitative statements. Within distinct evaluative criteria (agronomic traits, yield, processing qualities, and market value), participating

Fig. 1 A farmer-managed variety selection trial for triadic comparisons of technologies (tricot) in Honduras (*left*). Farmers evaluating an experimental trial for the accuracy assessment reported in this article (*right*)

farmers are asked to define each the best and the worst variety from within their set of three. Participants report their observations by answering two straightforward questions for each aspect that is being evaluated. For example, for yield, the questions would be "Which variety had the highest yield?" and "Which variety had the lowest yield?" This is an important reason for reducing the number of varieties to be tested by a single participant to three. If larger sets of varieties were to be ranked by each participant, more complicated questions would need to be asked. Straightforward questions are needed to be able to retrieve the information through telephone interviews, including automated calls.

2.2 Accuracy

We assess the quality of citizen science data produced by farmers by focusing on their accuracy. Accuracy consists of two components, reliability and validity (ISO 1994). The reliability of a method is its ability to produce repeated, consistent results. Validity refers to the closeness of a result or the mean of a large group of results to the actual value or accepted standard. The combined information about reliability and validity allows discussing the accuracy of a method.

The research method tested in this study is smallholder farmers' ranking of three different crop varieties according to observable plant characteristics. Reliability is expressed as the degree of internal agreement among observers about the ranking of varieties. Validity of the data is measured as the degree of agreement of farmers' observations with a ranking that was established by an agronomist who evaluated the same set of varieties, which we refer to as a "scientific ranking" (Sect. 2.3).

The comparison between farmers' rankings and a scientific ranking is not meant to question the overall validity of farmer *knowledge*, as this would imply problematic assumptions about these two forms of knowledge and their relative value (see Cleveland and Soleri (2007) for a discussion of the epistemological questions around comparing farmer and scientific knowledge). A key motivation for participatory research is to accommodate diverse viewpoints, and to tap into knowledge that is inaccessible, hard to interpret, or "invisible" to researchers. But this means that at the same time, it is important to establish whether different elements of farmer and scientific knowledge correspond to the same objective reality. A minimal degree of correspondence is an essential condition for a meaningful dialogue between farmers and scientists.

This study has therefore the limited goal of establishing the commensurability of farmers' and scientists' *observations* on the same phenomena (and not their knowledge as a whole). The point is to evaluate if farmers and scientists reach the same conclusions about varietal characteristics, as a starting point for subsequent farmer-scientist dialogue to make sense of these observations. We study varietal characteristics that are objectively observable rather than characteristics that involve a strong element of subjective assessment or preference (e.g., taste).

The tricot approach makes use of the trade-off between reliability and validity by placing emphasis on validity over reliability. As the Wisdom of Crowds principle suggests, a large sample of data may lead to a correct result even when individual data entries vary strongly (low reliability) as long as an unbiased aggregate measure can be calculated from the data (high validity). Tricot achieves external validity by placing crop varieties and other agricultural technologies directly in their target environment and by evaluating their performance in the eyes of the persons who will eventually adopt the technology or not. Independence of observations is ensured by not revealing the names of varieties or technologies and asking participants individually for their results. The Wisdom of Crowds requirement of having a diversity of viewpoints is fulfilled by inviting a diverse group of participants (women, men) to grow the varieties at many different plots, each one under slightly different crop management and environmental conditions.

2.3 Experimental design

At five sites in Honduras, small trials of three different varieties of common bean (*P. vulgaris* L.) were planted by collaborating farmers. These volunteers were smallholder farmers participating in tricot-style variety selection for common bean (see Van Etten et al. 2016). We assigned to each site a combination of three different varieties drawn from a total set of seven varieties. All varieties were phenotypically clearly distinct and uniform. Seeds were obtained from the bean breeding program at Zamorano Panamerican Agricultural University in Honduras. We randomized the assignment of combinations to sites and the order in which the varieties within each combination of three were ordered. The host farmers planted the three varieties of each combination at the same date, and each farmer managed their three varieties in the same way. They located the three varieties in each set directly next to each other in sub-plots with six rows of 8 m for each variety.

At five different points in the growing cycle, a total number of 35 smallholder farmers (18 women and 17 men) were asked to evaluate the three varieties at one of the sites (Fig. 1). In each session, groups of five to eight farmers participated. The selection of participants was determined by ongoing work of two local NGOs, and no additional criteria besides a balanced gender ratio were applied. The participants were first informed about the format of the exercise and that they would be asked to evaluate four agronomic traits: *Plant vigor*, *Plant architecture*, *Pest resistance*, and *Disease resistance*. In earlier participatory research, local farmers and breeders had established these traits as the most important pre-harvest selection criteria for bean varietal improvement (Steinke 2015), and they are common criteria in participatory variety selection for common bean (Asfaw et al. 2012).

Participants were then asked to take a few minutes to familiarize themselves with the three varieties planted, and focus on observable expressions of the traits. From the earlier research experiences, the farmers were acquainted with the concepts of Plant vigor—a merger of leaf area, leaf color, and physiological plant state (e.g., absence of drought stress symptoms)—and Plant architecture, for which farmers prefer non-trailing, upright-growing plants. But the enumerator also rephrased the exercise using local farmers' common wording, like "how well the foliage has developed" (for Plant vigor) and "how nicely the plant stands/grows" (for Plant architecture). For pest and disease resistance, participants were asked to acknowledge the presence or absence of attack symptoms, in order to identify different resistance capacity of varieties indirectly. The rationale behind observing the occurrence of biotic stressors as an inverse proxy for resistance requires the assumption that pest and disease pressure on the three trial varieties is equal, and the intensity of attack symptoms is thus determined largely by differences in genetic resistance. The questions asked were "which variety is (least/most) affected by (pests/diseases)?"

Except for the individual host farmers, participants had not seen the trials before. The importance of independent, individual assessments was emphasized when explaining the experiment to the participants, and participants were requested to refrain from exchanging their ideas about the varieties, in order to guarantee independent observations. The participants did largely remain silent during the evaluation.

After a few minutes of observing all four traits of the three varieties, the farmers were approached individually by the enumerator. The enumerator asked for their view on which was the best and the worst variety regarding each of the four criteria and recorded the answers. In each of the sessions, a local agricultural expert, in all cases an agronomist with much field experience working with common bean, also answered the same questions, and these assessments were taken as the respective scientific ranking for each of the different sites to measure the accuracy of farmers' observations against (Sect. 2.5).

Due to differences in planting dates and growing environments, the trials were in different development stages during the fieldwork period. This limited the observations that could be made in different sessions with farmers. In particular, pest and disease incidence cannot be evaluated before plants enter the reproductive phase (approximately 35 days after sowing), so these observations were only collected at two out of five sites.

2.4 Data preprocessing

For each plant characteristic, participating farmers indicated which they found to be the best and worst out of three varieties planted in the trial, coded A, B, and C. By inserting the implicit medium-ranked variety, every individual observation was converted into a ranking pattern, for example C > B > A. Incomplete observations and ties were removed from the data. Given the small number of observations per session, we decided to pool data from all sites by plant characteristic. For each site, farmer observations were recorded in relation to the expert's ranking order. At every site and for every evaluative criterion, the best variety according to the expert was coded variety X, the second-best variety Y, and the worst variety Z. This way, all valid farmer observations on one evaluative criterion could be converted in a standardized way to a permutation of X > Y > Z, the scientific ranking order. This way of data pooling assumes that there are no important differences between the sites in terms of the difficulty to discriminate between varieties. This is a reasonable assumption because at all sites, the local expert was able to rank the varieties for all plant characteristics in an unambiguous way (e.g., no ties between varieties); thus, any differences in rankings are mainly due to farmers' observation and interpretation ability.

2.5 Kendall's tau coefficient

To approach validity of observations, we quantified deviations of farmer rankings from the respective scientific ranking with Kendall's tau coefficient (τ), a measure of similarity between two rankings (Kendall 1938). The τ between two rankings is defined as follows:

$$\tau = \frac{C-D}{n(n-1)/2}$$

where C is the number of concordantly ranked item binaries (e.g., X > Y) between the two ranking lists, D is the number of discordantly ranked binaries, and n is the total number of binaries. τ may take values from −1 (completely reverse ranking) to 1 (identical ranking). In our case, the correct ranking pattern is always defined as X > Y > Z. In this case, a stated farmer observation of X > Z > Y or Y > X > Z gives $\tau = 0.33$, and Y > Z > X or Z > X > Y gives $\tau = -0.33$. Distributions of τ can be compared to the expected distribution of τ under a random null model. Under the null model, $\tau = 1$ is expected to occur in one out of six random rankings, $\tau = 0.33$ in two out of six, $\tau = -0.33$ in two out of six, and $\tau = -1$ in one out of six. To test whether there is an influence of gender on variety preferences or data quality, we performed Wilcoxon's signed rank test on the distributions of Kendall's tau coefficients of men's and women's observations for each of the four plant characteristics.

2.6 Mallows-Bradley-Terry model

For every plant characteristic, we fit a Mallows-Bradley-Terry (MBT) model (Mallows 1957; Tversky 1972) to the observed frequencies of the variety ranking patterns. Our criterion for validity was whether the MBT model was able to correctly distinguish the three varieties from each other at the $p < 0.05$ significance level. To reduce the risk of type I error due to multiple hypothesis testing, we performed p value corrections by the Holm-Bonferroni method (Holm 1979), a conservative method for controlling the family-wise error rate.

2.7 Kendall's *W*

We assessed reliability by determining the concordance between participants. We used Kendall's W to quantify the internal reliability for multiple dependent rankings (Kendall and Babington-Smith 1939). Kendall's W may take values ranging from 0, representing completely random results and no noticeable concordance among observers (rankers), to 1, meaning total agreement among all observers. We converted Kendall's W into verbal statements on agreement (from "very weak" to "unusually strong"), following the classification proposed by Schmidt (1997).

2.8 Simulations

Sample size choices will depend on trade-offs between research costs and data quality in different contexts. To inform such decisions, we created different scenarios with different numbers of varieties (n_{var}) and participants (n_{obs}). For each scenario, we determined the *discriminative ability*, defined as the number of varieties that can be statistically distinguished from the best variety ($p < 0.05$), as a simple heuristic.

We represent the observable performances of the varieties by a normally distributed variable, following a variation of Henrich and Boyd's (1998) simple model of environmental learning. We assume equal inter-variety intervals between varieties, and equal standard deviations (SD = 1). We estimated inter-variety interval values from the data by fitting the Thurstone-Mosteller case V (TM) model, which assumes that underlying parameters are normally distributed with an equal standard deviation of 1 (Mosteller 1951a, b). We chose the TM model for ease of interpretation because—like Henrich and Boyd's environmental learning model—the TM model uses Gaussian distributions, whereas the (Mallows-)Bradley-Terry model uses Gumbel distributions.

From the results of the TM model, we calculated the mean interval between trial varieties using the TM parameter estimates (mean of Y-X and Z-Y). This represents the mean pairwise performance difference among three varieties drawn randomly from a total pool of seven varieties. To obtain a representative inter-variety interval, we further divided the average X-Y-Z interval by 2, the mean number of intervals separating two varieties when three varieties are drawn out of a set of seven.

From the calculated performance intervals for all observed variety traits, we only retained the highest and lowest mean interval (Plant vigor and Disease resistance) for the simulations, thus testing one "easy" and one "challenging" plant characteristic. We generated 18 sets of modeled crop varieties, each containing $n_{var} \in \{3, 4, 5, \ldots, 20\}$ varieties. Also, we created six sets with different numbers of observers, $n_{obs} \in \{10, 20, 50, 100, 200, 500\}$. This resulted in 18 variety sets × 6 farmer sets × 2 different variety traits = 216 different scenarios.

We ran the simulations 1000 times for each of the 216 scenarios. For each run, we created a balanced experimental design. To simulate an individual participant's observation, we drew three varieties from the overall set of n_{var} varieties following the experimental design. For these three varieties, we then drew random numbers from their respective normal distributions. Subsequently, we compared these values to create a ranking. We repeated this for all n_{obs} participants in the set. We then ran the generalized Bradley-Terry-ε model on the resulting rankings (Firth 1993). This model will not break down if one variety wins or loses from all other varieties (unlike the classic BT model) and works with more than six

Table 1 Kendall's tau coefficient, standard deviation (SD), and Kendall's *W* of experimental farmer variety rankings

Variable	Frequency of observations with Kendall's tau coefficient (τ)				Mean τ	SD	Observers	Kendall's *W*
	$\tau = 1$	$\tau = 0.33$	$\tau = -0.33$	$\tau = -1$				
Plant vigor	64%	36%	0%	0%	0.76	0.32	22	0.676**
Plant architecture	54%	23%	19%	4%	0.51	0.60	26	0.280**
Pest resistance	46%	38%	15%	0%	0.54	0.48	13	0.337*
Disease resistance	27%	55%	9%	9%	0.33	0.57	11	0.174
Random null model	17%	33%	33%	17%	0	–	–	–

Percentages do not always add to 100 because of rounding. Significance values for the calculation of Kendall's *W* are as follows: *$p < 0.05$, **$p < 0.001$

varieties (unlike the MBT model). It is commonly used on ranking data and leads to consistent rankings (Jeon and Kim 2013). As a simple performance measure, for each of the 1000 runs, we determined the number of varieties that could be distinguished from the best variety at the $p < 0.05$ significance level, the discriminative ability. For each of the 216 scenarios, we calculated the median discriminative ability, as well as percentiles ($p = 5$ and $p = 95$).

2.9 Computational resources

For statistical analysis, we used the R programming language and environment (R Core Team 2016). We calculated Kendall's tau coefficient (Sect. 2.5) with the *Kendall* function of package *Kendall* (McLeod 2011). To fit the win counts for MBT models (Sect. 2.6) with the *glm* function (R Core Team 2016), we constructed paired comparison matrices with the *patt.design* function of package *prefmod* (Hatzinger and Dittrich 2012) and extracted *p* values with the *stars.pval* function of package *gtools* (Warnes et al. 2014). Kendall's *W* (Sect. 2.7) was calculated using the *kendall* function of package *irr* (Gamer et al. 2012). For the simulations (Sect. 2.8), we fit TM models with the *thurstone* function of package *eba* (Wickelmaier and Schmid 2004) and BT models using the functions *countsToBinomial* and *BTm* of package *BradleyTerry2* (Turner and Firth 2012). To speed up the simulations, we ran foreach loops, using the *doParallel* package (Calaway et al. 2015), and used the *plyr* package to reformat data (Wickham 2011).

3 Results and discussion

3.1 Accuracy of farmer-generated data

Table 1 presents the share of each τ value among all observations on each plant characteristic. In the case of Plant vigor, all observers fully or almost agreed with the scientific ranking. Observations on Plant architecture and Pest resistance are slightly less clear-cut, with a mean τ to the scientific ranking of about 0.5 each. Observations on Disease resistance are, on average, most divergent, with a mean τ to the scientific ranking of 0.33.

Wilcoxon's signed rank test on the τ values of men's and women's evaluations did not reveal a gender effect on observation validity for any of the plant traits at the $p < 0.05$ significance level (Table 2). The scientific literature provides evidence for gender-biased agricultural capacity, resulting from gendered household domains, such as the cultivation of different crops by women and men, gendered focus on different steps of food production and processing in the household, or contact to extension (Peterman et al. 2010; Quisumbing et al. 2014). Such gender differences may translate into different observation accuracies for different traits. In this study, however, all participants were currently engaged in cultivating bean. We would expect a stronger gender effect on agronomic knowledge in situations where the task division between women and men is more pronounced.

As can be seen in Table 1, correct observations with $\tau = 1$ were consistently more frequent than a random distribution would suggest, and, in return, incorrect observations with $\tau = -0.33$ or $\tau = -1$ were less frequent. Only for rankings on Plant architecture were observations with $\tau = 0.33$ less frequent than a random distribution would suggest. For Disease resistance, $\tau = 0.33$ has higher frequency than $\tau = 1$. Under the random null model, twice as many cases with $\tau = 0.33$ are expected than with $\tau = 1$, as two rankings are possible for $\tau = 0.33$ (Sect. 2.5), so this does not necessarily mean that

Table 2 Mean Kendall's tau coefficient (τ) and standard deviation (SD) of men's and women's observations on four plant traits, and *p* value of Wilcoxon's signed rank test between gender-disaggregated observations

Variable	Women			Men			*p* value
	Mean τ	SD	*n*	Mean τ	SD	*n*	
Plant vigor	0.74	0.33	13	0.78	0.32	9	0.841
Plant architecture	0.67	0.49	12	0.33	0.67	14	0.189
Pest resistance	0.33	0.50	7	0.78	0.32	6	0.140
Disease resistance	0.11	0.63	6	0.60	0.33	5	0.227

the consensus about Disease resistance does not converge to the scientific ranking. A more synthetic approach to determine validity is to use the MBT model.

Table 3 presents the results of MBT model estimation, including Holm-Bonferroni-adjusted p values. For all variables, the MBT model gives the correct ranking order; i.e., the estimate differences have the correct, negative sign, and |X-Z| > |X-Y|. For Plant vigor, the MBT model not only gives the correct order but also detects significant differences between all three varieties. For Plant architecture, all variety binaries but the best to the second-best varieties can be distinguished from each other. For Pest resistance, the expert-assessed best and worst varieties can be distinguished from each other. For Disease resistance, no variety can be distinguished from another at the $p < 0.05$ significance level. Nonetheless, we observe that (i) in all cases, the groups of observers converged on the same order as the agronomists (X > Y > Z) and (ii) except for Disease resistance, they were able to distinguish the best from the worst variety at the $p < 0.05$ significance level. This test was based on empirical data with a small number of observations, and in the next section, we explore the consequences of these findings with increased sample sizes.

We assessed Kendall's W as a measure of reliability for all traits (Table 1). For rankings on Plant vigor, Kendall's W is 0.676, a value indicating strong agreement among the observers. Rankings on Plant architecture achieve Kendall's W of 0.280, and rankings on Pest resistance achieve Kendall's W of 0.337, revealing weak agreement among observers in both cases. Rankings on Disease resistance result in Kendall's W of 0.174, which may be interpreted as very weak to weak agreement. Kendall's W was significantly higher than zero in all cases, except for Disease resistance. However, Disease resistance was the evaluative criterion for which we had the smallest sample size (n_{obs} = 11), giving it very small statistical power.

Depending on the trait, 77–100% of the observations match or nearly match the scientific ranking ($\tau = 1$ or $\tau = 0.33$), while only a 50% match would be expected if the rankings were completely random and contained no information. For all four traits, even with low numbers of observers, the MBT model ordered the varieties in the correct order, and for three traits, the model determined that the best and the worst variety performed significantly different from each other. So regardless of the varying levels of reliability in the data, our results were valid in all cases of our experiment.

Reliability, however, is only high for one variable, Plant vigor. This outcome relates to the expected difficulty of participants in observing the traits. Plant vigor can be assessed easily from a distance, and differences in leaf development and color intensity can be pronounced between crop varieties. Both Plant architecture and Pest resistance require some closer inspection of individual plants and leaves, which may also be somewhat more time-consuming. Lastly, the correct observation of diseases (or their absence), especially at early stages, demands more thorough scrutiny and background knowledge, including techniques of observation. Lack of training and awareness about diseases may be suggested as a reason leading to the relatively lowest validity, i.e., the highest degree of incorrect observations on Disease resistance. Our results concur with Bentley's (1989) reasoning that the ease of visual observation is an important determinant of the accuracy of farmers' observations and is therefore a main factor explaining the depth and level of concurrence of farmers' and formal scientific knowledge in different domains.

The relatively short time available for farmers' on-site evaluations in the experimental procedure we applied may also explain observed differences in accuracy to some extent.

Table 3 Results of Mallows-Bradley-Terry model estimation of farmers' variety rankings

Variable	Varieties		Estimate difference	Standard error	z value	p value (unadjusted)	p value (Holm-Bonferroni correction)
Plant vigor	X	Y	−0.895	0.293	−3.050	0.002**	0.005**
	Y	Z	−0.609	0.239	−2.543	0.011*	0.011*
	X	Z	−1.504	0.371	−4.049	5.152	0.000***
Plant architecture	X	Y	−0.204	0.154	−1.326	0.185	0.185
	Y	Z	−0.410	0.164	−2.498	0.012*	0.025*
	X	Z	−0.614	0.178	−3.449	0.001**	0.002**
Pest resistance	X	Y	−0.285	0.227	−1.252	0.211	0.211
	Y	Z	−0.429	0.240	−1.789	0.074	0.147
	X	Z	−0.713	0.270	−2.640	0.008**	0.025*
Disease resistance	X	Y	−0.150	0.226	−0.663	0.507	0.507
	Y	Z	−0.301	0.234	−1.283	0.199	0.399
	X	Z	−0.451	0.246	−1.832	0.067	0.201

Varieties X, Y, and Z represent the expert-assessed best, second-, and third-best varieties at each experimental site, respectively

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Although participants were not being rushed in our experiments, we expect that farmers in future tricot-style on-farm variety trials would get a better insight into pest and disease resistance, as farmers will be able to observe the plants on multiple occasions throughout a growth cycle, and follow the evolution of pests and diseases over time.

Rankings that differ from the scientific ranking do not necessarily reflect a random error due to a lack of observational or diagnostic capacity, but may indicate that some participants had a semantic understanding of the concept that is divergent from the expert's understanding. For example, the concept of good plant architecture may vary among farmers, so the observers giving reverse or near-reverse rankings ($\tau = -1$ or $\tau = -0.33$) may actually have assessed correctly according to their own criteria. It may be possible to detect the presence of disagreements statistically (cf. Mueller and Veinott 2008). The detection of substantial disagreements could be used as a data quality diagnostic tool in future applications.

This study only focused on pre-harvest plant characteristics. The tricot methodology can also be employed for assessing harvest and post-harvest variety characteristics, such as yield, cooking time, and processing or storage qualities. While the findings on observation accuracy can perhaps be generalized to other vegetative plant characteristics, the experimental process described here should be repeated in order to assess the appropriateness of the tricot method for producing findings about post-harvest variables.

3.2 Discriminative ability simulations

The mean inter-variety performance interval from the TM model estimation was highest for Plant vigor (1.03) and lowest for Disease resistance (0.37). Only these values (after dividing by 2, as explained in Sect. 2.8) were used in the simulations. Figure 2 shows the median discriminative abilities, i.e., the numbers of varieties that could be distinguished from the best variety at the $p < 0.05$ significance level, as well as the respective number of varieties that could not be distinguished. In the simulation results, we observe three patterns.

Our first observation is that discriminative ability increases with an increasing number of observers. For example, the discriminative ability for Plant vigor with $n_{var} = 12$ goes up from four varieties ($n_{obs} = 10$) to ten varieties ($n_{obs} = 500$). When more observers are engaged, pairwise combinations of two varieties are replicated more often, which in turn leads to more accurate parameter estimates and a higher discriminative ability.

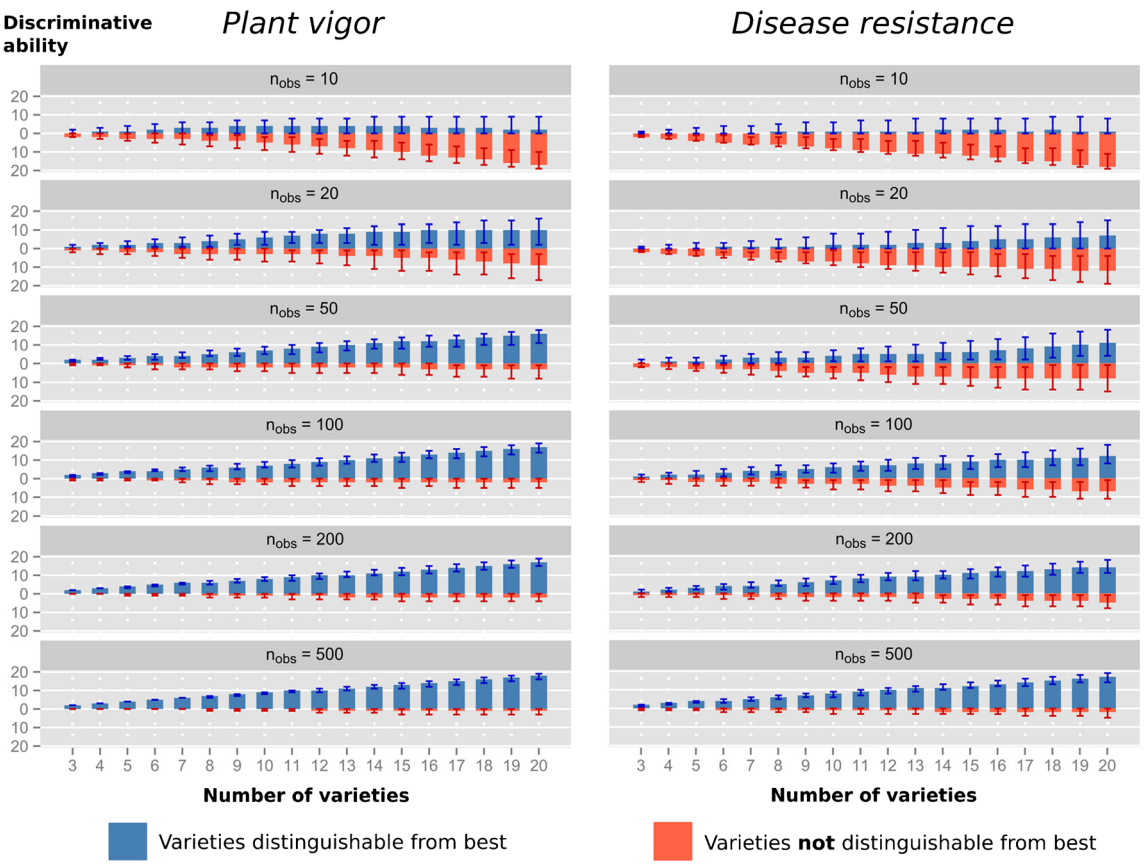

Fig. 2 Simulated discriminative abilities of various research scenarios for tricot. n_{obs} = number of observers. *Bars* represent the number of varieties that can and cannot be distinguished from the best variety, for the "easy observation" trait *Plant vigor*, and the "challenging observation" trait *Disease resistance*. Median values and percentiles (5, 95) of 1000 runs are shown

Secondly, we observe that discriminative ability increases when adding more varieties to the evaluation. It does so within successive sets of n_{var} values, within which every $n_{var} + 1$ leads to an increase in the number of distinguishable varieties, while the number of non-distinguishable varieties remains stable. For example, for Plant vigor and $n_{obs} = 50$, within the range of $n_{var} \in \{7, \ldots, 14\}$, every additional variety in the roster leads to an increase in discriminative ability. But because with every set of n_{var} values, the number of varieties that *cannot* be distinguished from the best increases by one, the relative share of distinguishable varieties among all evaluated varieties decreases overall with increasing n_{var}. For Plant vigor and $n_{obs} = 50$, for example, the share of distinguishable varieties decreases from 100% at $n_{var} = 3$ to 84% at $n_{var} = 20$. This decrease was expected; when more varieties are included in the scenario, while keeping the observer number constant, the number of evaluations of each pairwise combination of two varieties decreases. The share of distinguishable varieties may be used as a measure of efficiency of experimental design in the tricot approach.

Thirdly, for the same scenario, i.e., the same combination of n_{var} and n_{obs}, our simulations predict higher discriminative ability for Plant vigor than for Disease resistance, with few exceptions of no difference at $n_{obs} = 10$ and $n_{obs} = 500$. This was to be expected, as the inter-variety intervals (relative to the standard deviation, set to 1) for Plant vigor are higher than for Disease resistance. Hence, the discriminative ability of a given number of observers will depend on the expected reliability for the tested trait, which itself depends on the ease of visual observation. For the same scenario, the discriminative ability is usually lower for Disease resistance than for Plant vigor due to the lower reliability of observations. Engaging more participants can compensate this effect. For example, our simulations indicate that the discriminative ability reached for Plant vigor with $n_{var} = 12$, and 50 participants would take 200 participants for Disease resistance.

How these results translate into sample size decisions will depend on the objective of variety selection. For example, a staged selection could be done, first focusing on the more easily observable characteristics. For such a first stage, only the reliability of observations on these easy traits would need to be taken into account. Also, the reliability of the observations can be increased by more training on disease recognition and other relatively challenging traits. In practical applications of the tricot approach, maximum or near-maximum discriminative ability may not be necessary. For example, the ability to identify a 50% share of varieties that perform worse than the best one may be the main aim of certain applications, e.g., to identify promising varieties at an initial on-farm screening step.

For the correct interpretation of our simulation results, it is important to note that our model assumes an idealized situation, where observable performance intervals between varieties are regularly spaced. In real life, such clear-cut differences between crop varieties are not to be expected, so discriminative ability is likely to be smaller than it is in our simulations. The selected set of varieties may include varieties that are similar for a number of traits. When the performance of varieties is virtually equal, discriminative ability may be affected. But at the same time, distinguishing between tiny differences in variety performance on farms is of limited practical relevance. Another important limitation of the current study is that it has taken into account only one potential source of error, that is, farmers' observations. Other sources of error can include experimental errors, or cases in which seeds and codes have been mixed up at some stage of the process. Also, attrition rates have not been taken into account, e.g., participants who drop out from the tricot experiment before successfully ending the trial, due to external factors or a lack of interest. To determine minimum sample sizes in real experiments, these additional factors need to be taken into account.

Furthermore, the indicated sample sizes suppose that the results are valid across the entire group of participants, which is true only in the absence of strong genotype-by-environment interactions or preferences influenced by gender, culture, or socio-economic status. Accounting for environmental gradients or doing a gender-differentiated analysis is possible, for example, by using BT models with "recursive partitioning," a method to distinguish groups of observers with different preference profiles (Strobl et al. 2011). In this case, researchers will need to revise the participant numbers upwards in order to attain reasonable results. They may use a simulation approach similar to the one presented here to assess how many participants are needed.

4 Conclusions

Our results show that in the triadic comparisons of technologies (tricot) citizen science methodology, the relatively low reliability of individual results does not undermine the accuracy of the findings when a sufficiently large group of farmers participates. Low reliability of farmer observations is no hindrance to obtaining statistically significant and relevant results. Our results show that, in aggregate, the observations contain sufficient information. Larger numbers of observations are expected to lead to statistical modeling results that distinguish between more varieties. In other words, the Wisdom of Crowds principle applies in this context: sufficiently large numbers of observers can compensate low reliability of observations as long as there is good validity, i.e., when the consensus of this large group converges on the correct answer. This means that scaling on-farm agricultural research by a crowdsourcing methodology is feasible.

Variation in farmers' observations, leading to decreased reliability, is caused not only by incorrect observations, e.g.,

due to the challenging evaluation of some plant traits, but also by possibly divergent views on varietal quality indicators among observers. Such differing reference systems may stem, e.g., not only from local variation in environmental pressures but also from group-specific, e.g., gendered preferences. While low reliability from either source can be balanced by engaging higher numbers of observers to achieve significant distinction of varieties, results from tricot-style research necessarily reflect an averaged approach to farmers' understandings of tested traits, as well as their possibly varying preferences. In ongoing research, we are currently testing statistical methods that treat variation as information and that lead to alternative models, disaggregating results, e.g., along cut-points on environmental gradients.

For the varietal characteristics tested in this study, it was possible to reproduce scientific judgments through crowdsourcing farmer observations. Whether the same approach can be used to tap into farmer knowledge that is embedded in context and is inaccessible to scientists, and thereby elicit technology rankings that cannot be performed by conventional methods, remains to be tested. Our simulation results show that the order of magnitude of the group of participants required to achieve accurate results is reasonable given the logistical abilities of many organizations. Assuming an attrition rate of 20% or less, we estimate that in evaluations of sets of about 10–12 varieties, groups of 150–200 participants are likely to be sufficient to produce meaningful findings. But these results need to be revisited when more studies using the tricot approach become available. Some investment in training farmers to observe certain traits can pay off if this reduces the error significantly. Results may improve over time when farmers repeat participation over a number of crop cycles.

The possibility of citizen science via triadic comparisons of technologies opens interesting perspectives for agricultural science, beyond crop variety research. By testing technologies across environmental or socio-economic gradients, the acceptability of sets of research products can be estimated in a robust and cost-efficient way, informing the targeting of these products to certain environments and types of farms. Compared with other farmer-participatory research methodologies, adopting a "hands-off" citizen science approach reduces requirements for logistics, farmer training, field visits, and physical assets per participant. With limited resources, research organizations may reach both higher numbers and a higher diversity of farming households for the specification of technologies under development, like unreleased crop varieties. Maybe more importantly, tricot-style research can integrate new research products continuously. With every crop cycle, for example, the worst-performing fraction of the materials (varieties, lines, clones, landraces, etc.) may be exchanged with new ones. This way, through iterative research cycles, technology specification may improve, and individual participant farmers' experimentation may benefit from knowledge generated by the Wisdom of Crowds.

Recent approaches to agricultural extension have stressed the need to link stakeholders for knowledge exchange and social learning, as well as the need to facilitate autonomous experimentation with innovations (Desclaux et al. 2008; Schut et al. 2016). Steinke and van Etten (2016) also encourage researchers employing the tricot methodology to bring together farmer citizen scientists in workshops. Yet, the benefit of the citizen science approach is that it poses a low entry threshold to those farm households who are regularly excluded from both traditional and modern approaches to extension and participatory research due to remoteness, time and labor constraints, or social conflict. Through tricot, participation in agricultural research and extension may be feasible with very low additional effort and little modification to regular farm-life activities. In addition, as observations are performed individually, under real-life farm conditions, and trait-by-trait along the crop cycle, selection will incorporate information about the variation among farmers and environments. Farmer groups working with collective plots tend to mask much of this variation (cf. Misiko 2013). Through reductions in staff time and logistics, we expect higher cost-efficiency of the approach, which we currently quantify in ongoing research. We also test the possibility of detecting the influence of environment (climate and soil) and other variables on farmer observations, and the effect of the tricot approach on farmer learning, which is an important goal of participatory research.

To researchers interested in implementing the tricot approach, we recommend to plan their research based on a preparatory order-of-magnitude study following a similar protocol as the one presented here, as levels of discriminative ability in practice are likely to vary. A preparatory study could also detect farmers' semantic disagreement about concepts. If such disagreement is found, it can be countered by ensuring consensus about the concepts through a good explanation or by capturing the subjective element in the evaluations in a different way. Learning and exchange of experiences should iteratively help to improve the design and execution of tricot trials.

Acknowledgments This work was implemented by Bioversity International as part of the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS). The views expressed in this document cannot be taken to reflect the official opinions of CGIAR or Future Earth. Crowdsourcing crop improvement: Evidence base and outscaling model. The authors are very grateful to Dr. Juan Carlos Rosas and staff at Zamorano, as well as the field facilitators of the Honduran NGOs PRR and FIPAH for the support on site. We would like to thank Dr. Helen Ogden (Warwick) for an important pointer on statistics and Vincent Johnson for editing the manuscript. We thank five anonymous reviewers,

whose contributions substantially improved the manuscript. Above all, we are grateful to the participating farmers for their commitment, time, and effort.

References

Asfaw A, Almekinders CJM, Blair MW, Struik PC (2012) Participatory approach in common bean (Phaseolus vulgaris L.) breeding for drought tolerance for southern Ethiopia. Plant Breeding 131:125–145. doi:10.1111/j.1439-0523.2011.01921.x

Bentley JW (1989) What farmers don't know can't help them: the strengths and weaknesses of indigenous technical knowledge in Honduras. Agric Hum Values Summer:25–31. doi:10.1007/BF02217666

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs. I. The method of paired comparisons. Biometrika 39:324–345. doi:10.1093/biomet/39.3-4.324

Calaway R, Weston S, Tenenbaum D (2015) doParallel: foreach parallel adaptor for the 'parallel' package. Available at: https://cran.r-project.org/web/packages/doParallel/index.html

Classen L, Humphries S, Fitzsimmons J, Kaaria S, Jiménez J, Sierra F, Gallardo O (2008) Opening participatory spaces for the most marginal: learning from collection action in the Honduran hillsides. World Dev 36:2402–2420. doi:10.1016/j.worlddev.2008.04.007

Cleveland D, Soleri D (2007) Farmer knowledge and scientist knowledge in sustainable agricultural development: ontology, epistemology and praxis. In: Sillitoe P (ed) Local science vs global science: approaches to indigenous knowledge in international development. Bergahn, New York, pp 209–230

Coe R (2002) Analyzing ranking and rating data from participatory on-farm trials. In: Bellon MR, Reeves J (eds) Quantitative analysis of data from participatory methods in plant breeding. CIMMYT, Mexico, D.F., pp 44–65

Desclaux D, Nolot JM, Chiffoleau Y, Gozé E, Leclerc C (2008) Changes in the concept of genotype × environment interactions to fit agriculture diversification and decentralized participatory plant breeding: pluridisciplinary point of view. Euphytica 163:533–546. doi:10.1007/s10681-008-9717-2

Dickinson JL, Shirk J, Bonter D, Bonney R, Crain RL, Martin J, Phillips T, Purcell K (2012) The current state of citizen science as a tool for ecological research and public engagement. Front Ecol Environ 10:291–297. doi:10.1890/110236

Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80:27–38. doi:10.2307/2336755

Fürnkranz J, Hüllermeier E (eds) (2010) Preference learning. Springer, Berlin and Heidelberg

Gamer M, Lemon J, Fellow I. Singh P (2012) irr: various coefficients of interrater reliability and agreement. R package version 0.84, available at: http://cran.r-project.org/package=irr

Hand E (2010) People power. Nature 466:685–687. doi:10.1038/466685a

Hatzinger R, Dittrich R (2012) prefmod: an R package for modeling preferences based on paired comparisons, rankings, or ratings. J Stat Softw 48:1–31. doi:10.18637/jss.v048.i10

Humphries S, Rosas JC, Gómez M, Jiménez J, Sierra F, Gallardo O, Avila C, Barahona M (2015) Synergies at the interface of farmer–scientist partnerships: agricultural innovation through participatory research and plant breeding in Honduras. Agr Food Sec 4:1–17. doi:10.1186/s40066-015-004

Hellin J, Bellon MR, Badstue L, Dixon J, La Rovere R (2008) Increasing the impacts of participatory research. Expl Agric 44:81–95. doi:10.1017/S0014479707005935

Henrich J, Boyd R (1998) The evolution of conformist transmission and the emergence of between-group differences. Evol Hum Behav 19:215–224. doi:10.1016/S1090-5138(98)00018-X

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70

ISO (1994) Accuracy (trueness and precision) of measurement methods and results—part 1: general principles and definitions. BS ISO 5725-1. International Organization for Standardization, Geneva

Jeon JJ, Kim Y (2013) Revisiting the Bradley-Terry model and its application to information retrieval. J Korean Dat Inf Sci 24:1098–1099. doi:10.7465/jkdi.2013.24.5.1089

Kendall MG (1938) A new measure of rank correlation. Biometrika 30:81–93. doi:10.2307/2332226

Kendall MG, Babington-Smith B (1939) The problem of m rankings. Ann Math Stat 10:275–287

Kremen C, Ullman KS, Thorp RW (2011) Evaluating the quality of citizen-scientist data on pollinator communities. Conserv Biol 25:607–617. doi:10.1111/j.1523-1739.2011.01657.x

Lilja N, Dixon J, Eade D (eds) (2013) Participatory research and gender analysis: new approaches. Routledge, London

Martin GJ (2004) Ethnobotany: a methods manual. Earthscan, London

Mallows CL (1957) Non-null ranking models. I. Biometrika 44:114–130. doi:10.2307/2333244

McLeod AI (2011) Kendall: Kendall rank correlation and Mann-Kendall trend test. R package version 2.2. http://CRAN.R-project.org/package=Kendall

Misiko M (2013) Dilemma in participatory selection of varieties. Agric Syst 119:35–42. doi:10.1016/j.agsy.2013.04.004

Morris ML, Bellon MR (2004) Participatory plant breeding research: opportunities and challenges for the international crop improvement system. Euphytica 136:21–35. doi:10.1023/B:EUPH.0000019509.37769.b1

Mosteller F (1951a) Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. Psychometrika 16:3–9

Mosteller F (1951b) Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. Psychometrika 16:207–218

Mueller ST, Veinott ES (2008) Cultural mixture modeling: a method for identifying cultural consensus. ARA Technol Rev 4:39–45

Peterman A, Behrman J, Quisumbing A (2010) A review of empirical evidence on gender differences in nonland agricultural inputs, technology, and services in developing countries. In: Quisumbing et al. (eds) Gender in agriculture. Closing the knowledge gap. Food and Agriculture Organization of the United Nations, Washington, D.C, pp 145–186

Quisumbing AR, Meinzen-Dick R, Raney TL, Croppenstedt A, Behrman JA, Peterman A (2014) Closing the knowledge gap on gender in agriculture. In: Quisumbing et al (eds) Gender in agriculture. Closing the knowledge gap. Food and Agriculture Organization of the United Nations, Washington, D.C, pp 3–27

R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Schmidt RC (1997) Managing Delphi surveys using nonparametric statistical techniques. Decis Sci 28:763–774. doi:10.1111/j.1540-5915.1997.tb01330.x

Schut M, Klerckx L, Sartas M, Lamers D, Campbell MMC, Ogbonna I, Kaushik P, Atta-Krah K, Leeuwis C (2016) Innovation platforms: experiences with their institutional embedding in agricultural

development. Expl Agric 52:537–561. doi:10.1017/S001447971500023X

Snapp S, Heong KL (2003) Scaling up and out. In: Pound B, Snapp S, McDougall C, Braun A (eds) Managing natural resources for sustainable livelihoods. London and Sterling, Uniting Science and Participation, International Development Research Center/ Earthscan, pp 67–87

Spielman DJ, Ekboir J, Davis K (2009) The art and science of innovation systems inquiry: applications to sub-Saharan African agriculture. Technol Soc 31:399–405

Steinke J (2015) Citizen science with resource-poor farmers as a new approach to climate adaptation and food security: evidence from Honduras. Humboldt University Berlin, Master's thesis

Steinke J, van Etten J (2016) Farmer experimentation for climate adaptation with triadic comparisons of technologies (tricot). A methodological guide. Bioversity International, Rome

Strobl C, Wickelmaier F, Zeileis A (2011) Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. J Educ Behav Stat 36:135–153. doi:10.3102/1076998609359791

Surowiecki J (2005) The Wisdom of Crowds. Anchor Books, New York

Turner H, Firth D (2012) Bradley-Terry models in R: the BradleyTerry2 package. J Stat Softw 48:1–21. doi:10.18637/jss.v048.i09

Tversky A (1972) Elimination by aspects: a theory of choice. Psychol Rev 79:281–299. doi:10.1037/h0032955

Van Etten J (2011) Crowdsourcing crop improvement in sub-Saharan Africa: a proposal for a scalable and inclusive approach to food security. IDS Bull 42:102–110. doi:10.1111/j.1759-5436.2011.00240.x

Van Etten J, Beza E, Calderer L, van Duijvendijk K, Fadda C, Fantahun B, Kidane YG, van den Gevel J, Gupta A, Mengistu DK, Kiambi D, Mathur PN, Mercado L, Mittra S, Mollel M, Rosas JC, Steinke J, Suchini JG, Zimmerer KS (2016) First experiences with a novel farmer citizen science approach: crowdsourcing participatory variety selection through on-farm triadic comparisons of technologies (tricot). Expl Agric. doi:10.1017/S0014479716000739

Warnes GR, Bolker B, Lumley T (2014) gtools: various R programming tools. R package version 3.4.1. http://CRAN.R-project.org/package=gtools

Wickelmaier F, Schmid C (2004) A Matlab function to estimate choice model parameters from paired-comparison data. Behav Res Methods Instrum Comput 36:29–40. doi:10.3758/BF03195547

Wickham H (2011) The split-apply-combine strategy for data analysis. J Stat Softw 40:1–29. doi:10.18637/jss.v040.i01

Witcombe JR, Joshi A, Joshi KD, Sthapit BR (1996) Farmer participatory crop improvement. I. Varietal selection and breeding methods and their impact on biodiversity. Expl Agric 32:445–460. doi:10.1017/S0014479700001526