




RESEARCH ARTICLE

Reproducibility and external validity of on-farm experimental research in Africa

Hanna Kool¹ , Jens A. Andersson^{1,2}  and Ken E. Giller^{1,*} 

¹Plant Production Systems Group, Wageningen University, P.O. Box 430, 6700 AK Wageningen, the Netherlands and

²International Maize and Wheat Improvement Center (CIMMYT), United Nations Avenue, Gigiri, Box 104, 00621 Village Market, Nairobi, Kenya

*Corresponding author. Email: ken.giller@wur.nl

(Received 3 September 2019; revised 10 March 2020; accepted 12 June 2020; first published online 17 July 2020)

Abstract

Agronomists have increasingly conducted experiments on-farm, in an attempt to increase the wider applicability (external validity) of their experimental findings and their relevance for agricultural development. This review assesses the way in which on-farm experimental studies address the scope or generalisability of their findings when based on a limited number of farms. A central question is how on-farm studies define the environment or research population in which the on-farm trial findings are valid, or are valuable for. Such an assessment is, of course, conditional on the (internal) validity of the experimental findings. We therefore first analyse how authors of on-farm experimental studies describe the factors that may shape experimental outcomes. As agronomic experiments often use ‘yield’ as dependent variable to assess treatment effects, we developed a procedure to score studies on their descriptions of yield-determining factors. Although experimental validity principally rests upon the reproducibility of the experiment and its findings, we found that on the basis of the information provided in published on-farm experimental studies, it is often difficult or impossible to reproduce the experimental design. Nutrient management, weed management and crop information are best described, whereas land preparation, field history and management of pests and water are rarely described. Further, on-farm experimental studies often compare treatments to a ‘farmer practice’ reference or control treatment which is assumed to be widely and uniformly practiced and known to the reader. The wider applicability or external validity is often poorly addressed in the reviewed studies. Most do not explicitly define the research population and/or environment in which (they expect) the experimental findings to work. Academic textbooks on agronomic experimentation are remarkably silent on both the internal and external validity of on-farm experimentation. We therefore argue for more systematic investigations and descriptions of the research population and settings to which on-farm experimental studies seek to generalise their findings.

Keywords: Reproducibility crisis; Internal validity; On-farm experiments; Control treatments; Farmer practice

Introduction

Agriculture is experimental by nature as it is geared towards improving farming results (Maat, 2011; Richards, 1985). Yet in agricultural research, experiments are conducted by agronomists in two distinct locations: on research stations and in farmers’ fields. Research stations represent a highly controlled experimental environment, where single or multiple treatments and their interactions can be investigated holding other variables constant. In farmers’ fields, researchers have much less control over the experimental situation, which is embedded in a much wider and more variable bio-physical and socio-economic environment.

© The Author(s), 2020. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

An increased focus on on-farm experimentation has been driven by two distinct but inter-related lines of investigation. First, in Africa, research stations were often located on good soils in environments considered optimal for breeding new varieties of important export crops (Vanlauwe *et al.*, 2019). Recognition that smallholder farming took place on less fertile soils in a range of agroecologies stimulated on-farm research on fertilizer responses in important staple crops such as maize and beans (e.g. Anderson, 1974; Scaife, 1968). The finding that the implicit assumption of consistency and reproducibility of experimental findings between research stations and farmers' fields was untenable (Gomez and Gomez, 1984; Waddington *et al.*, 2007), increasingly pushed development-oriented agronomists to evaluate new technologies and practices in on-farm experiments. For instance, the huge diversity in smallholder farmers' resource endowments and use of organic manures has major effects on crop response to fertilizers (Giller *et al.*, 2011; Zingore *et al.*, 2007). Hence, in order to increase the wider applicability or external validity of experimental studies and to understand the underlying factors influencing the impacts of new technologies, agronomists have increasingly conducted experiments on-farm (Vanlauwe *et al.*, 2019). Second, on-farm experiments became particularly popular following the rise of Farming Systems Research (FSR) and Farmer Participatory Research (FPR) in the 1980s (Byerlee *et al.*, 1982; Chambers and Jiggins, 1987; Collinson, 2000). Both of these lines of investigation sought to make agricultural research more relevant for farmer circumstances, and its products more widely adoptable by (resource-poor) smallholder farmers in the global south. With that in mind, we would expect that on-farm experiments pay more attention to external validity (e.g. the wider application) than on-station experiments. The quest for enhancing the wider relevance of agronomic experiments that drove the increase of on-farm experimentation appears to be highly compatible with the current drive for 'impact at scale' in development-oriented research. It is therefore not surprising that on-farm experiments have remained a central research method in development-oriented agronomy, albeit that different approaches to on-farm experimentation can be distinguished.

A first approach – 'common on-farm experimentation' – evolved from the drive towards increased on-farm experimentation. Moving agronomic experimentation from research stations onto farmers' fields implied less controlled and more heterogeneous experimental situations. In response, agronomists conducting on-farm experiments with multiple replicates usually work on a number of farms. However, resource constraints and the logistics of on-farm experimentation often limit the number of experiments – as we found in this study, commonly to less than 20 farms (see Figure 3). Consequently, increasing the external validity of on-farm experimental research forefronts the choice of location, research population and experimental treatments, as these determine outcomes. However strong the experimental design and implementation, the results are useless if the on-farm experiment is conducted with the 'wrong type' of farmer, ending up in the 'wrong type' of field, or when the control and other experimental treatments have no bearing on the farming context of that location.

A second 'stratified-experimentation' approach aims to represent a range of environments. It builds on a stratified sampling strategy, where locations of on-farm experiments are purposely selected along different agroecological zones (Kaizzi *et al.*, 2012; Trutmann and Graf, 1993), landscape positions (Ebanyat *et al.*, 2010) or soil fertility gradients (Vanlauwe *et al.*, 2006; Zingore *et al.*, 2007).

A third more recently developed approach to on-farm experimentation builds on larger numbers of experimental observations specifically designed to understand the variability in response to treatments. These 'experimentation-at-scale' studies are sometimes referred to as development-to-research (Giller *et al.*, 2013) or research-in-development approaches (Coe *et al.*, 2014) as they are often conducted in the context of large-scale development projects. Such approaches do not seek to control variability in experimental conditions. Rather, they embrace the variability in experimental situations and responses and seek to understand how contextual factors shape experimental outcomes (Vanlauwe *et al.*, 2019). Characterized by simple experimental designs and treatment variables, often with farms serving as a single replicate (block), these approaches generally start with defining the research population or extrapolation domain and then focus on defining an appropriate random sampling frame. Such studies specifically focus on identifying factors that shape the variability in experimental outcomes among farms (Bielders and Gérard, 2015; Ronner *et al.*, 2016).

The three distinguished approaches to on-farm experimentation represent different strategies for generalisation from experimental findings. While arguably the experimentation-at-scale approach does not seek to generalise findings beyond a pre-defined research population, defining the scope or environments in which the experimental findings are valid or valuable for remains a major task for any agronomist conducting on-farm experiments. Although the vast majority of studies selected for this review appear to fall into the category of ‘common on-farm experimentation’, our conclusions are applicable to all three approaches.

Before being able to assess how on-farm experiments are formulated in relation to the wider environments in which their results should be applicable, it is first necessary to assess the internal validity of experimental designs. As it is impossible to assess in detail the internal validity of a large number of studies within the scope of this article, we focus on this by looking at the possible factors influencing experimental outcomes and whether study descriptions adequately cover such influences. As we cannot claim to have analysed an exhaustive dataset of on-farm studies in Africa, we first review the main topics and geographical distribution of the on-farm experimental studies included, to provide the reader with an understanding of the context of our analysis.

Although research topics and objectives of on-farm experimental studies in Africa differ widely, they very often – and are similar to on-station trials in this respect – use ‘yield’ as the most important dependent variable to assess the effects of experimental treatments. We therefore begin our analysis with an assessment of the descriptions of the factors that determine yield that are provided in the reviewed experimental studies. After all, experiments are only the most rigorous way to establish causal relations, if both their design and findings can be reproduced. We develop a procedure to score experimental studies on their descriptions of these different yield-determining factors; these have been well established in the literature on theoretical production ecology (Tittonell and Giller, 2013; van Ittersum and Rabbinge, 1997).

After reviewing the *internal* validity of on-farm experimental studies through a focus on their reproducibility, we shift attention towards the ways in which such studies define the wider applicability of the experimental results, that is, the *external* validity of the research. External validity asks the question of generalisability: To what populations, settings, treatment variables and measurement variables can this effect be generalised? We assess the external validity by focusing on the studies’ descriptions of the wider research population and research setting, as well as the generalisability of treatment and measurement variables (Campbell and Stanley, 1963). We conclude by making suggestions for a more systematic investigation and description of the research population and settings in on-farm experimental studies.

Material and Methods

Selection of on-farm papers in sub-Saharan Africa

Our analyses of on-farm experiments in Africa are based on a literature search conducted with the Web of Science (v.5.22.3) on October 28, 2016. We used the Web of Science™ Core Collection database, where we entered the following keywords in the topic field: ((experiment* OR trial*) AND farm* AND Africa¹). We selected a timespan from 1986 to 2015 and the following citation indexes were included: Science Citation Index Expanded – 1945-present, Social Science Citation Index – 1956-present and Emerging Sources Citation Index – 2015-present. This search resulted in 1005 papers. The search was refined by selecting the Research Areas: (AGRICULTURE, ENVIRONMENTAL SCIENCES, ECOLOGY, PLANT SCIENCES, ENTOMOLOGY, WATER RESOURCES and FORESTRY), reducing the number of papers to 767. We excluded review papers, leaving us 744 papers.

¹A search using country names instead of ‘Africa’ doubled the total number of papers. As this alternative search procedure did not alter the overall distribution of studies over the different countries, we assume that the findings from our analyses also apply to this larger ‘all-country’-based dataset.

The 744 papers were each examined and excluded when the experimental study was: 1) not conducted in sub-Saharan Africa (SSA); 2) not conducted on-farm, but on-station, or the location was unclear; 3) not a field experiment, but a pot, greenhouse or laboratory experiment; 4) a natural experiment (in which the assignment of treatments is not done by the researcher); 5) the paper was not about agronomy, but wildlife, nature, aquaculture, health or livestock; 6) the experimental data were published elsewhere and 7) not building on the on-farm situation, for example, as the experiment assessed the effect of different rainfall intensities by using a rainfall simulator. This resulted in 172 papers which were used in our detailed analyses (see for more details, Appendix I and Supplementary Materials).

Analysis of the number of experimental locations

Although the issue of generalisation is relevant to any on-farm experimental study, our review revealed that the ‘common on-farm experimentation’ approach is indeed the most common, and therefore our findings pertain especially to this category. To distinguish ‘experimentation-at-scale’ and ‘stratified experimentation’ from the majority of ‘common on-farm experimentation’ approaches – we first identified the number of experimental locations per study. The number of experimental locations is the number of fields or farms where the full experiment was conducted. When studies conducted multiple experiments, we only reported the experiment with the largest number of experimental locations. Although studies are often multi-seasonal (generally covering 2 to 3 seasons), the number of experimental locations reported refers to the largest number in one growing season. The reason for this is that, first, it is not always clear whether repeated experiments were actually repeated on the exact same location or in different locations. Second, the duration of some experiments extends beyond one season, for example, when rotation or residual effects of previous crops are studied.

Analysis of reproducibility

Since the vast majority of the analysed studies used yield as (the most important) dependent variable, we assessed the capture of yield determining factors in the studies’ descriptions of the experimental treatments and their results. Such information is not only important for the assessment of the study’s findings, but also critical to its reproduction. Building on the common understanding that crop yields are determined by genotype (G), environment (E) and management (M) interactions, van Ittersum and Rabbinge (1997) distinguish yield-defining, yield-limiting and yield-reducing factors. Yield-defining factors are those that, at optimum supply of all inputs, determine the potential production level, such as the plant characteristics, the temperature and solar radiation at the geographical location. Yield-limiting factors refer to the abiotic factors – such as water and nutrients – that limit plants to develop to their full potential, whereas growth or yield-reducing factors such as weeds, pests and diseases and pollutants impede plant growth. In order to assess the experimental studies’ capture of these different factors in their descriptions of the experimental treatments (in the M&M and Results sections), these factors were translated into seven categories of yield-determining variables (Table 1).

The seven categories are in the order as operations are conducted throughout the cropping season starting with the category ‘field history’, which provides insight in the initial status of field. This is followed by crop information and land preparation at the start of the growing season, while during the growing season, nutrient, water, pest and weed management become increasingly important.

The category ‘crop information’ reflects the *yield-defining factors*, with the variables ‘varieties’, ‘planting date’, ‘plant spacing’ and ‘mono-/intercrop, crop rotation’. ‘Varieties’ can be found in all three groups (yield-defining, -limiting, and -reducing factors). Next to how varieties define the

Table 1. Seven categories containing 23 yield-determining variables were defined and grouped as yield-defining, yield-limiting and yield-reducing factors. (G), (E) and (M) refer to genotype, environment and management interactions, respectively. The variables were used in the scoring procedure to assess the reproducibility of studies

| Categories | Yield-defining factors | Yield-limiting factors | Yield-reducing factors |
|----------------------------|---|---|---|
| <i>Field history</i> | | Past crops grown (M) Past soil/nutrient/pest management (M) Soil chemical and physical properties (E) | Past crops grown (M) Past soil/nutrient/pest management (M) |
| <i>Crop information</i> | Varieties (G) Planting date (M) Plant spacing (M) Mono-/intercrop, crop rotation (M) | Varieties (G) | Varieties (G) |
| <i>Land preparation</i> | | Tillage method (M) Timing (M) Tillage depth (M) | Tillage method (M) Timing (M) Tillage depth (M) |
| <i>Nutrient management</i> | | Type nutrient (M) Application method (M) Quantity (M) Timing (M) | |
| <i>Water management</i> | | Rainfall quantity (E) Rainfall distribution (E) Irrigation (M) Irrigation frequency/quantity (M) Water harvesting techniques (M) | |
| <i>Pest management</i> | | | Pest management method (M) Frequency/quantity of pest management (M) |
| <i>Weed management</i> | | | Weed management method (M) Frequency/quantity of weed management (M) |

crops' growth rate, morphology and plant architecture, other variety characteristics, such as its water and nutrient use efficiency, pest and weed resistance/tolerance, affect the impact of water and nutrient shortages and pests and diseases on the crop's yield (Tittonell and Giller, 2013).

The categories 'nutrient management' and 'water management' refer to the *yield-limiting factors*, while 'pest and weed management' reflect the *yield-reducing factors*. The categories 'field history' and 'land preparation' include both *yield-limiting factors* as well as *yield-reducing factors*. Land preparation can affect nutrient and water availability, but can also be used to control weeds. For the category 'field history', 'past pest management' is a yield-reducing factor, whereas 'soil chemical and physical properties' and 'past soil and fertilizer management' are yield-limiting factors, as they provide insight in the nutrient availability at the start of the experiment.

Scoring procedure

Each on-farm study's description of the used experimental treatments was scored on the basis of 23 variables, resulting in a maximum 'reproducibility' score of 23 for studies that provide information on all 23 variables. If any information was given about a defined variable, the study would receive a point. Some variables contain multiple descriptions, such as the variables 'Past soil/nutrient/pest management', 'Frequency/quantity of pest management', 'Irrigation frequency/quantity' and 'Mono-/intercrop, crop rotation'. A point was assigned when studies described one (or more) of these aspects of this variable.

The scoring procedure focused on the information provided on these variables, not on whether the practice was done. For example, if no nutrients were applied in an experiment, and this fact

was explicitly stated, the study received all points for the category ‘Nutrient management’. For the variable ‘Soil chemical and physical properties’, a point was assigned, when soil tests were conducted before the start of the experiment. No point was assigned to studies conducting a soil test at a later stage of the experiment. For the variables ‘rainfall quantity’ and ‘rainfall distribution’, studies were scored when they described these variables for the corresponding growing season. When only the average amount of rainfall over many years was given or the rainfall distribution was simply described as bimodal or unimodal, no points were assigned.

Assessment of external validity

As the wider applicability of a study is primarily defined by its research question, an analysis of the external validity of numerous studies is necessarily indirect and of a general nature. Not being able to directly assess the external validity of the on-farm studies – as this would also require independent field research – we focus on how the scope of the study or research population is defined. It appeared that the reviewed studies rarely explicitly address in what environmental conditions or for what farmer populations the experimental results are deemed valid or relevant. We therefore also included other elements from which information regarding the wider applicability of the experimental findings or the research population can be gleaned: (1) general descriptions of the research setting or environmental conditions; (2) selection criteria used for field, farmer and site selection; (3) descriptions of the sample field, farms and farmers and (4) the definition of the ‘farmer practice’ treatment, if such a treatment is included in the experimental design.

Results

On-farm experimental research in Africa

Although the literature search we conducted is not exhaustive, covering only publications on on-farm experiments included in the Web of Science™ Core Collection for the period 1986–2015, Figure 1 suggests a clear trend. Following the emergence of FSR and FPR approaches in the 1980s, on-farm experimental studies have increasingly made it into high-ranking agronomy journals. But while reflecting the overall trend, the annual figures in Figure 1. need to be treated with caution, as both the number of publications and the information on publications have also increased over the years; with the inclusion of more data in publication records, such as abstracts or more keywords, the chance of new publications to be harvested in a keyword search increases.

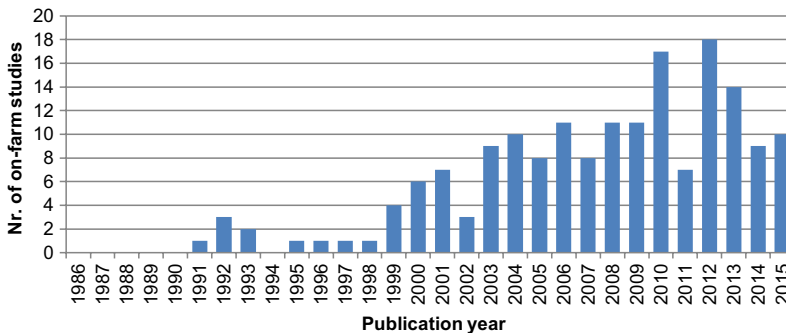


Figure 1. Number of on-farm experimental studies in sub-Saharan Africa published in the period 1986–2015 (Web of Science Core collection).

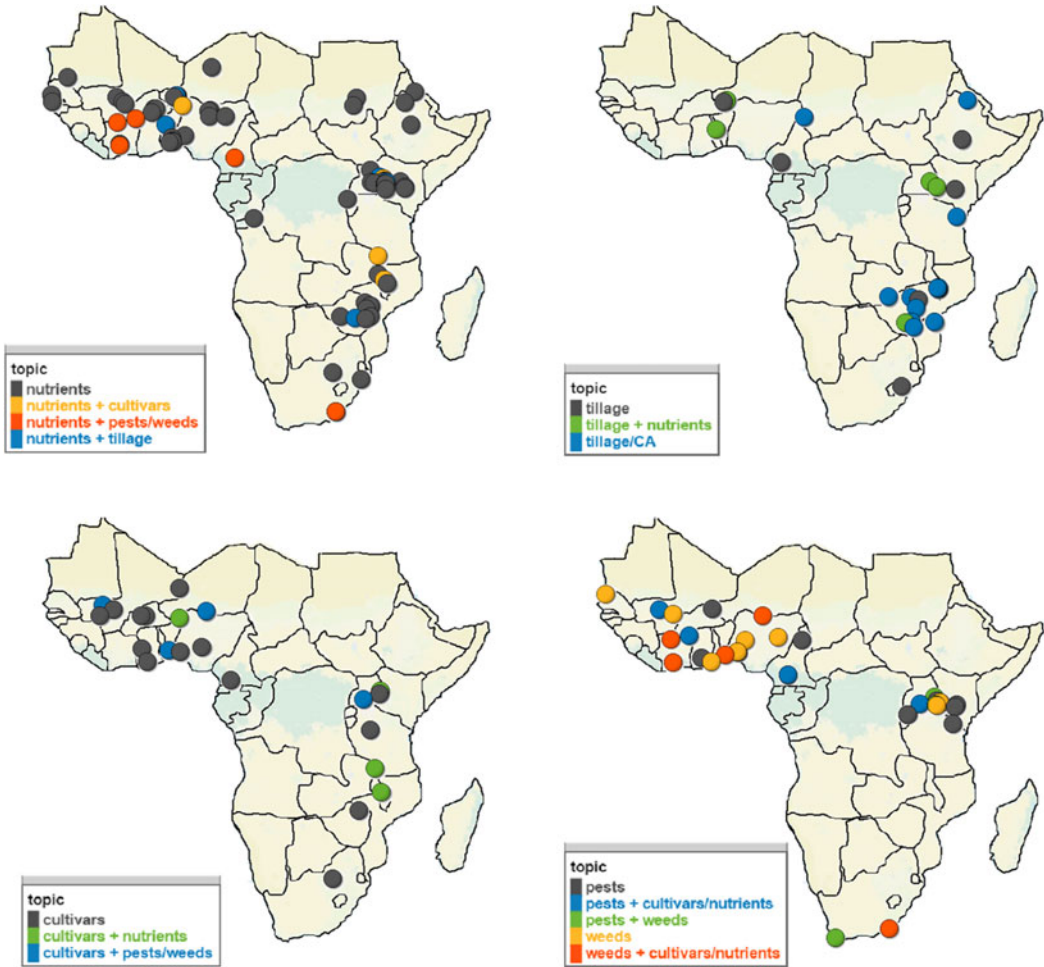


Figure 2. Locations of on-farm experimental studies in SSA (n = 138) by their main topic of the experiment: Nutrients (top left, n = 75), Tillage (top right, n = 26), Pest and Weeds (below left, n = 36) and Cultivars (below right, n = 22). Each dot represents one on-farm study. 34 studies had other topics and were not plotted. Some studies combine topics, they appear on more than one map.

Geographical distribution of on-farm experiments by topic

The on-farm experimental studies considered in this review are distributed across SSA, except for a band stretching from Namibia and Botswana towards Chad and Sudan, where only few published on-farm studies were conducted. Hotspots can be found in East and West Africa, notably in countries with substantial presence of international agricultural research institutes of the CGIAR. To gain insight in the geographical distribution of the on-farm experimental studies, each study was first categorised based by its main topic, as reflected in the title and abstract. The study locations of the four most common topics: (a) nutrients, (b) tillage, (c) pests and weeds and (d) cultivars are plotted in the different maps of Figure 2. (Studies that address combinations of these topics are indicated in different colours on the same map.)

On-farm experimental studies in SSA (n = 172) appear to focus predominantly on the effects of nutrients on crop growth (n = 75), studying the effects of mineral fertilizers, composts, manure, crop residues or combinations thereof. These ‘nutrient’ studies are similarly geographically distributed as all studies. By contrast, ‘pests and weeds’ studies (n = 36) appear to be relatively

concentrated in West-Africa, Kenya and Uganda, while ‘tillage’ studies (n = 26) were concentrated in Southern Africa (Zimbabwe, Malawi, Mozambique and Zambia) and mainly concerned Conservation Agriculture. Studies with the topic ‘cultivars’ (n = 22) are similarly distributed as the ‘nutrient’ studies.

Treatments in on-farm experiments

The treatments used in the on-farm experiments (Table 2) may not reflect the main topic of the paper. For example, fertilizer rates may be varied between treatments even when ‘nutrients’ are not the main topic of the paper. Variable fertilizer rates or types appeared to be the most common difference between treatments, as 54% of the experiments investigated different rates or sources of organic and/or chemical fertilizer.

Number of experimental locations of on-farm studies

Figure 3 presents the number of experimental locations per on-farm study. Most studies (78%) have less than 20 locations. More than half of the studies (52%) are conducted in 1 to 10 different locations, of which a large part is only in 1 or 2 locations. Only 3% of the studies are large scale with 100 up to 1000 locations. These findings suggest that most of the included studies are location specific and follow the ‘common on-farm experimentation’ approach.

Table 2. What varies between the different treatments in on-farm studies? Note that the sum of these percentages is above 100%, as experimental treatments may vary in multiple ways

| Treatment | Percentage of studies |
|---|-----------------------|
| Mineral fertilizer | 47 |
| Pest and weed management | 22 |
| Organic fertilizer | 19 |
| Legume | 17 |
| Cultivars | 16 |
| Tillage/Conservation Agriculture | 16 |
| Plant spacing and timing | 7 |
| Water | 3 |
| Other variables, such as pruning, fallows and crop rotation | 12 |
| SUM | 158 |

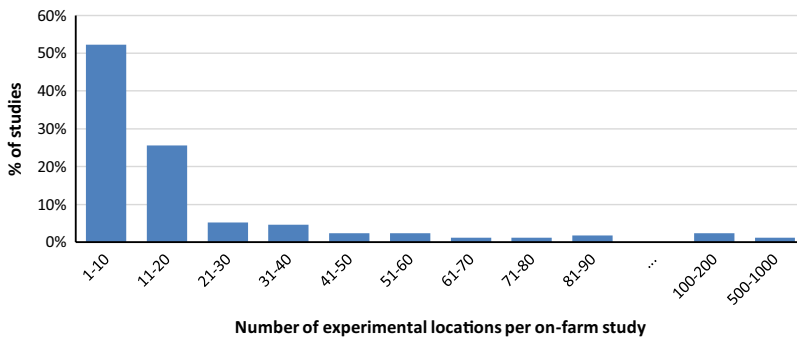


Figure 3. The number of experimental locations per on-farm experimental study. A quarter of the studies (24%) were conducted in only 1 or 2 locations.

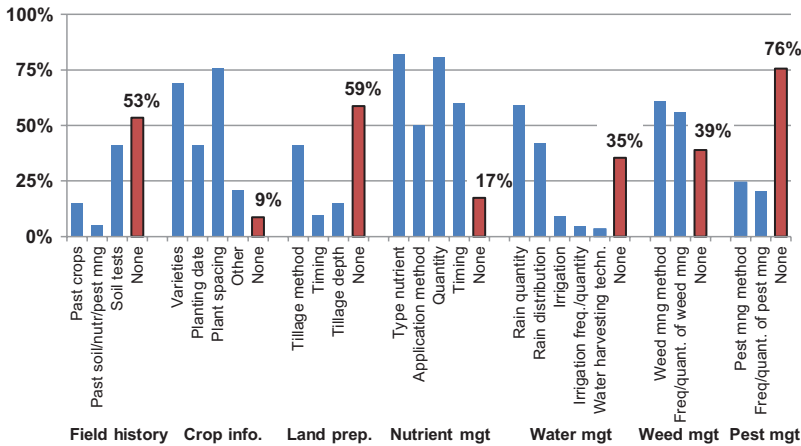


Figure 4. Percentage of studies which describe the variables of seven categories (from left to right: field history, crop information, land preparation, nutrient management, water management, weed and pest management). ‘None’ indicates the number of studies that mentioned nothing about that category.

Internal validity – reproducibility

Key to any assessment of an experimental study’s internal validity is the question whether the experimental results can be reproduced. In on-farm experimentation, this is, of course, problematic as there is only limited control over the experimental situation. Not all experimental conditions can easily be reproduced. Yet, an assessment of any experimental study’s internal validity and findings is conditional on the possibility of reproducing its design and set-up and to know about key experimental conditions. To assess the reproducibility of on-farm experiments, we scored the 172 on-farm experimental studies on their description of 23 yield-defining, yield-limiting and yield-reducing factors. We considered an experiment to be reproducible when the paper includes descriptions of these factors. This does not mean that the *findings* are reproducible given that some variables such as rainfall will differ between seasons. Information on such variables will, of course, be useful for interpretation of the results. A plotting of the scores of all studies mimics a normal distribution and ranges between 1 and 18, with an average score of 8.9 out of 23 (=39%) and a median score of 9. This suggests that few published studies provide the information needed to repeat the on-farm experiment.

Figure 4 presents scores per category. Overall, nutrient management, weed management and crop information are best described, with an average score of 68, 58 and 52%, respectively.

The red bars ‘None’ at each category indicate the percentage of studies which did not describe any of the variables of that category. The percentages of studies failing to describe a category *at all* is as follows: pest management (76%), land preparation (59%), field history (53%), weed management (39%), water management (35%), nutrient management (17%) and crop information (9%). These results suggest that not all categories of yield-determining factors get equal attention. Below we zoom in on each of the categories.

Field history

Less than half of the studies (47%) provide descriptions of the experimental field’s history. Elaborate reporting is rare; only three studies (2%) conducted both a soil test *and* provided descriptions of past crops, past soil, nutrient and pest management practices. Most studies that describe the experimental field’s history (41%) conducted soil analysis before the experiment started. Only 15% of studies mention the crops grown in the previous season(s), and only 5% described past soil, nutrient or pest management.

Since the effect of nutrients on crop growth is the most studied topic in on-farm experiments in Africa, these findings are perhaps not surprising. Apparently, agronomists are more interested in the initial nutrient status of the experimental field than in the possible causes of that nutrient status or its representativeness. Yet, as Falconnier *et al.* (2016) found in on-farm experiments in southern Mali, the previous crop, its management and nutrient carry-over from the previous season is often a major determinant of yield variability in farmer's fields.

In addition to insight into the (determinants of the) nutrient status of the experimental field, information on previous crops grown can have an important influence on pest and weed prevalence. For example, past crops might have been hosts of pests or parasitic weeds, which in turn may affect the current weed infestation. Yet only 5 out of 36 'Pest & Weed' papers provided information concerning previous crops grown on the experimental field. Even more striking is that none of the 'Pests & Weeds' studies described past soil, nutrient or pest management practices.

Crop information

The vast majority of the papers (91%) provided information on the crop in the experiment, although this was often far from complete. Only 10% of the studies described all four factors. Most studies provide information on the plant spacing (76%) and the variety used (69%) but fewer mention the planting date (41%). Information on crop rotation, inter- or mono-cropping was often absent, with only 21% of studies describing one (or more) of these aspects.

Land preparation

Apart from studies focusing on tillage, most studies (59%) do not describe how the land was prepared. Of the studies that do mention tillage methods (41%), only about a third also described the timing of land preparation and the tillage depth. Only 6% of the studies describes all three factors.

Nutrient management

As the effect of nutrients on crop growth is the most studied relation in on-farm experiments, it is not surprising that nutrient management is one of the best described practices. Overall 44% of the papers described all four nutrient management related factors. Yet, 17% of the studies do not state anything about it. Most studies (around 80%) described the type and the quantity of fertilizer used. Of those studies, only 3/4 also mentioned the timing of fertilizer application and 3/5 described the method of application.

Water management

Fifty nine and 42% of the studies described total rainfall and its distribution, respectively. Some studies simply provided a yearly average of rainfall or described the rainfall distribution as bimodal or unimodal, without describing how rainfall was distributed in that specific growing season. Only 9% of the studies mentioned the use of irrigation, and only half of those studies stated something about the quantity or frequency of irrigation. A mere 3% of the studies mentioned anything about water harvesting techniques.

Pest and weed management

Pest management was least described. Only 61 and 24% of the studies mentioned anything about weed and pest management, respectively. Most of these studies also mentioned the frequency of weeding or the quantity of herbicide (56%) or pesticide (20%) applied. Several papers indicated that the crop was affected by a pest, yet omitted to describe any pest or disease management. For example, Khan *et al.* (2006) assessed stem borer infestation without indicating pest management practices. Manu-Aduening *et al.* (2006) evaluated different cassava cultivars, based on criteria such as pest resistance and weed suppression; however, the authors did not describe whether and how pests and weeds were managed.

Can on-farm experiments be reproduced?

Our overall conclusion from this analysis of reproducibility is that in general insufficient information is provided to be able to repeat an experiment. Usually, experimental treatments are better described than other aspects of crop management, which are also important to ensure reproducibility. In most papers (91%), all treatments are equally (poorly or well) described. Some papers (8%) do not describe all treatments to the same extent, and the description of the control treatment was less complete compared with the other treatments.

A number of studies give rather unclear and vague descriptions of experimental practices. For instance, '[fertilizer was] applied at plant growing stage' (Ratnadass *et al.*, 2008), 'weeding was carried out when necessary' (Worou *et al.*, 2013) or 'weeds were regularly controlled' (Mucheru-Muna *et al.*, 2014). Some studies refer solely to 'common farmer practice' when describing practices, which means that authors unjustifiably assume that readers know what is commonly done by farmers in the area of research. For example, 'Planting, fertilizer application, disease control, hilling and weeding were all done by the farmer groups using their common practice' (Gildemacher *et al.*, 2011), 'manuring, planting, weeding and pest control were undertaken by the farm household following their normal farm management practice' (Ncube *et al.*, 2007), 'Farmers planted at their usual density' (Ndjeunga and Bationo, 2005) and 'Land preparation prior to sowing was as practiced by farmers in the area' (Tulema *et al.*, 2007). In none of these examples do the authors provide details on how this was done. Such uninformative descriptions provide insufficient information to make the experiments reproducible.

External validity of on-farm experiments

Addressing the external validity of on-farm experimental studies revolves around describing and delineating the research population – the farmers who have similar characteristics, and who farm under similar bio-physical conditions. The research population is first and foremost defined by the study's research question ([a] in Figure 5).

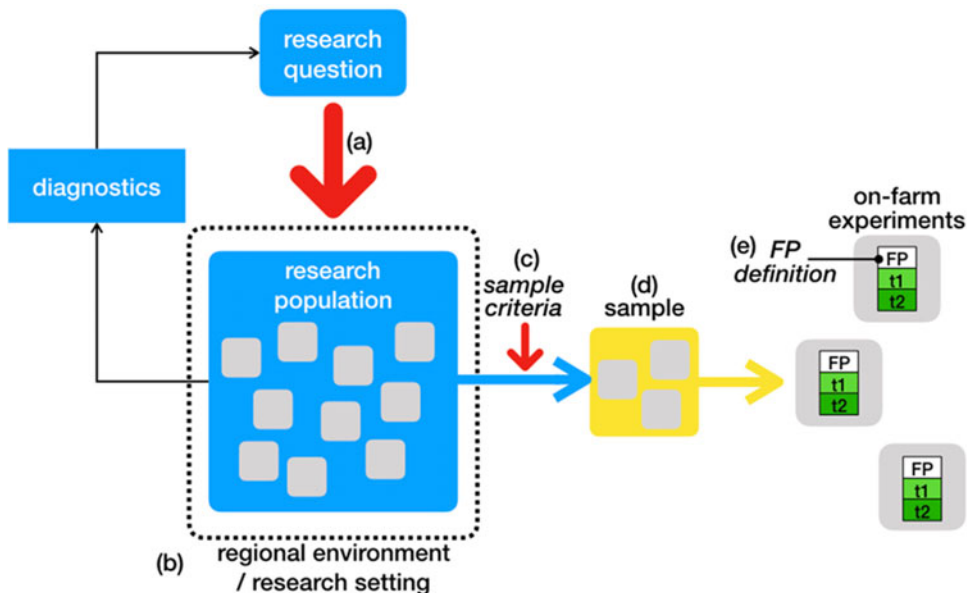


Figure 5. The research population is defined by the research question (a). In the reviewed studies the research population is referred to by: describing the regional environment (b), sample criteria (c), sample characteristics (d) and including a farmer practice (FP) treatment in the experimental design (e).

None of the reviewed studies explicitly defined the research population in which the experimental findings are applicable. However, the wider research population is usually referred to in indirect ways and can be gleaned from descriptions of:

- (1) the research setting or the region's characteristics – (b) in Figure 5;
- (2) the selection criteria used when selecting farms, farmers or fields for on-farm experimentation, or the characteristics of these selected farms – (c) and (d) in Figure 5;
- (3) the 'farmer practice' treatment, if the experiment includes such a treatment – (e) in Figure 5.

In the heyday of FSR, farming system diagnostics were seen as means to understand (and define) the research population and as key to the formulation of a research question (Mutsaers *et al.*, 1997; Stroud, 1993; Tripp, 1982). Such diagnostics appear to be out of fashion: only few of the reviewed studies reported on a pre-experimental diagnostic phase (Bucheyeki *et al.*, 2010; Manu-Aduening *et al.*, 2006; Tulema *et al.*, 2007).

Becker and Johnson (2001a)'s study in upland rice systems in Ivory Coast is one of few studies that provides insight into the regional importance of their on-farm experimentation. Studying the effect of weeding and fertilizer application, they conducted experiments in four different agro-ecological zones, for which they described not only the rice-based production systems but also specified the area size of the agro-ecological zones, as well as the share of rice growing area within these zones. This information provides a crucial insight into the extrapolation potential of the experimental results. Surprisingly, most on-farm studies do not report on the size of the research population.

Description of the study region's characteristics

Several studies describe the regional environment, such as the major crops grown in the area, agro-ecological zones, rainfall patterns, elevation and/or the main farming systems (Abdalla *et al.*, 2015; Franke *et al.*, 2010; Ojiem *et al.*, 2014; Roothaert *et al.*, 2003; Twomlow *et al.*, 2010). Socio-cultural aspects are sometimes also described. For example, Morse *et al.* (2009) elaborate on the ethnic groups and the languages spoken in the research area. Although such characteristics might give insight into the studied research population, their relevance for the extrapolation of experimental results is generally not made explicit. As a result, on-farm studies' descriptions of the regional environment or research setting generally provide little insight into the wider applicability of the experimental findings.

Criteria used for the selection of farms, farmers or fields

Criteria used for field, farm or farmer selection constitute another indirect way of finding out about the research population and the study's wider applicability. However, only a quarter of the reviewed studies (26%) elaborates on criteria used for selecting trial-hosting farmers/farms/fields. Studies that do elaborate on such criteria use widely diverging strategies, including both bio-physical and socio-economic criteria. For instance, where Becker and Johnson (2001b) selected sites representing different agro-ecological zones for lowland rice-growing (see above), Ojiem *et al.* (2007) selected different types of fields within farms, 'to represent high, medium and low soil fertility conditions'. Nyakudya *et al.* (2014) selected 'fields with contour ridges on slopes >4% with potential to generate large quantities of runoff. (...) Such fields were representative of most farmers' fields in the study area'. Rather than bio-physical characteristics, Twomlow *et al.* (2010) selected potential trial-hosting farming households based on their socio-economic status – 'households with limited cash income, female-headed households [and] households with high dependency ratio e.g. high numbers of children, orphans, handicapped, terminally ill and the elderly'.

Baudron *et al.* (2012) used tillage technology as an indicator of socio-economic differences, selecting both hand-hoe farmers and animal-drawn plough-owning farmers for inclusion in their on-farm experiments. A ‘sample of farmers was selected to represent these different farmer types’. Sometimes selection criteria are more vaguely described. For instance, ‘farmers possessing suitable fields were approached’ (Brocke *et al.*, 2010), or ‘fields . . . representative for cassava production in the area’ (Pypers *et al.*, 2012). Such descriptions do not contribute to our understanding of the external validity of the study.

While selection criteria potentially provide insight into an on-farm’s study external validity or wider applicability, descriptions are not necessarily informative. For instance, trial farm selection based on the accessibility of the site (Sanou *et al.*, 2014), ‘the ability [of farmers] to meet cost of land preparation, fertiliser, labour as well as evidence of record keeping’ (Fanadzo *et al.*, 2010), the willingness of farmers (Kearney *et al.*, 2012; Sanou *et al.*, 2014), their interest in the technology (Brocke *et al.*, 2010) or farmers being ‘socially well-integrated’ (Misiko *et al.*, 2008) provides few clues regarding the wider applicability of the experimental results. Although such selection criteria do not have a clear link with the research question, they may affect the study’s external validity, as they may reduce the size of the initially intended research population. For example, are experimental results of ‘accessible sites’ also applicable in (or relevant for) less accessible sites? Similarly, does the selection of farmers capable of record keeping unwittingly limit the research population to better educated, and possibly better-off, farmers? Many studies lack a reflection on what research population the selected farms/farmers represent, and how the selection criteria used to select experimental farms may further limit the wider applicability of the study.

References to ‘farmer practice’ treatments and their implications for external validity

As on-farm experimentation in the context of development-oriented research usually aims to improve upon existing farming practice, it is not surprising that on-farm experimental designs often include a ‘farmer practice’ reference treatment. We observed that 29% of the reviewed studies included ‘farmer practice’ treatments in the experimental design and predominantly, this was the control treatment. This research practice can be interpreted as a means to increase the study’s external validity – treatment effects are directly measured against a reference treatment that is used beyond the experimental plot. Enabling a comparison with existing practices, a ‘farmer practice’ treatment may be used to justify the promotion of the new, better performing technologies or management practices, tested in other experimental treatments. However, the incorporation of a ‘farmer practice’ treatment also introduces further concerns regarding the study’s wider applicability.

First, the ‘farmer practice’ treatment is often not clearly described (Becker and Johnson, 2001b; Pandey *et al.*, 2001; Rockström *et al.*, 2009; Thierfelder *et al.*, 2015; Yamoah *et al.*, 2011), which hampers an assessment of the external validity of the experimental findings. For instance, Nyamangara *et al.* (2014) studied the effect of conservation agriculture on maize yields in Zimbabwe, yet used a poorly described ‘conventional treatment’ as a ‘farmer practice’ control treatment. While for the promoted treatments, ‘planting basins’ and ‘ripper tillage’ information is given about the tillage depth and spacing and the nutrient application (quantity and type), this information is not provided for the ‘farmer practice’.

Second, many studies assume the ‘farmer practice’ to be uniform and do not recognise diversity among farmers and within farms (Fox and Rockström, 2000; Khan *et al.*, 2006; Lamers *et al.*, 2015; Ndjeunga and Bationo, 2005; Snapp *et al.*, 2002). A clear example is Rockström *et al.* (2009), who studied conservation agriculture in Kenya, Tanzania, Zambia and Ethiopia. The control treatment was defined as ‘conventional treatment’ or ‘conventional tillage’, which included only ‘animal drawn mouldboard ploughs’ and ‘pitting using hand-hoes’. Differences in tillage equipment – types of ploughs and hoes used – and tillage practices – number of passes, frequency, tillage depth, etc. – are not considered. Consequently, the studied minimum-tillage technique may as well be what

(some) farmers already practice. More importantly, however, the researcher's homogenising construct of a 'farmer practice' renders the assessment of the study's external validity problematic; which farmers, if any, actually practice the researcher-defined 'farmer practice'?

Whether the standardised 'farmer practice' treatment in on-farm experiments is actually common practice in the research population is hardly reflected upon. However, a few studies make use of non-standardised 'farmer practice' treatments (c.f. Franke *et al.*, 2010; Krupnik *et al.*, 2012; Twomlow *et al.*, 2010). For instance, Franke *et al.* (2010) used a non-standardised 'farmer practice' treatment in which farmers were free to grow their own varieties, implement their own fertilization strategy, crop choice, field management, etc. These diverse farmer practices and their results were recorded and compared with a baseline study to assess whether the participating farmers were better-off, farming on more fertile soils and/or using more inputs than the farmers included in the baseline. Interestingly, Franke *et al.* (2010) found that farmers copied management methods from some of the experimental treatments, and sometimes competed with the researcher-managed plots. Thus, they found that the mere presence of an on-farm experiment may influence the study's wider impact and relevance. Obviously, the use of a non-standardised 'farmer practice' results in more uncontrolled variation within the experiment, but any observed treatment effects are likely to have a greater external validity. Even non-standardised farmer-practice treatments are not without their problems, as the study of Franke *et al.* (2010) exemplifies.

Discussion

On-farm experiments and reproducibility: a crisis in Agronomy?

The recent 'reproducibility crisis' in a number of experimental sciences, notably social psychology and medicine (Baker, 2016; Pashler and Wagenmakers, 2012), has foregrounded the use of experimental method in contemporary scientific knowledge production. The failure to reproduce experimental findings has often been related to the academic environment in which contemporary research takes place. Amidst increased competition for research funds, mounting pressure to publish, a selection bias towards positive results in publishing etc., researchers may resort to questionable research practices – such as eliminating outliers, inappropriate rounding of *p* values, deciding to select more data after a significance check, failing to report (inter)dependent measures, etc (John *et al.*, 2012). Cases of academic fraud, such as plagiarism and data fabrication, have also been linked to this highly competitive academic environment (Baker, 2016), while an over-stretched peer-review system may be increasingly incapable of filtering out questionable research practices.

While the debate on reproducibility has mostly focused on the internal validity of experimental research (the correctness of the findings), relatively little attention has been paid to the reporting of experimental design, or the wider environment or population in which the experimental findings are also valid and relevant for. Such a focus is particularly relevant for an experimental science such as agronomy in which experimental conditions are highly location specific and difficult to control. Good descriptions of the experimental environment are not only important as they may shape experimental outcomes and the study's reproducibility, but also because they provide insight into the wider applicability of the experimental findings. The demand for research results that have wide applicability is high, and arguably, increasing in an era of impact-oriented agricultural research for development. This study therefore focused on how agronomists describe the environmental conditions of their on-farm experimentation.

Our review finds poor and little systematic description of experimental conditions, notably of yield-determining factors, the predominant dependent variable in on-farm experiments. As a consequence, it is not only difficult (or impossible) to repeat an on-farm experiment on the basis of the information provided in the scientific publication in which its results were presented. The lack

of systematic description often also prevents proper scrutiny of the presented experimental findings. Whether these findings should be labelled a ‘reproducibility crisis’ is a moot point. Since the results of on-farm experimental studies are seldom questioned, such a crisis does not seem to be experienced among agronomists.

External validity

The lack of consideration of the external validity of on-farm experimental studies is arguably, the most disturbing finding of this study, albeit not a new observation. Two decades ago, Mutsaers *et al.* (1997) already highlighted that many on-farm studies do not explicitly define a research population. References to the research population are often indirect and present an incomplete picture. Few studies justify or reflect on the selection of sites and treatments in view of the on-farm experiment’s external validity. Apparently, the wider applicability of on-farm experimental work is not considered a required topic to make experimental work publishable, neither by authors nor reviewers. This is remarkable, particularly in an era of result-oriented research funding and demands for ‘impact at scale’ in development-oriented agronomy.

We do not believe that a single guideline can be formulated, which can be followed as a recipe for addressing the external validity of agronomic experiments. The key conditions or characteristics of a research population depend on the research question, which are related to both bio-physical and socio-economic factors. Some characteristics might be relevant for one study, but not for another. Yet, in order to retain and increase the relevance of the experimental method in agronomy, we propose that experimental studies should minimally address: (1) under what conditions the treatment effect occurs and (2) for whom the experimental findings are relevant.

In addition, we advocate for more transparency about the selection criteria used for research site, farmer and field selection and a reflection on the possible implications of such selection for external validity. Especially studies using a ‘farmer practice’ treatment should critically reflect on whether this treatment and its performance is representative of the research population. As on-farm studies are strongly defined by the social context in which they are conducted (de Roo *et al.*, 2019; Andersson *et al.*, 2019), it may be useful for agronomists to (re-)define the research population also *ex-post*.

Agronomic textbooks and on-farm experiments

Our finding that on-farm experimental studies are characterised by poor and little systematic description of experimental conditions appears to reflect the guidelines in agronomic textbooks. For instance, while discussions of the experimental method in these textbooks acknowledge the importance of documenting site information and trial management, they stress the recording of such information in view of analysing and interpreting experimental results (CIMMYT, 1988; Patel *et al.*, 2004; Stroud, 1993; Tripp, 1982), not in order to enable the experiment to be repeated. Not surprisingly, agronomic textbooks do not explicitly consider the concept of reproducibility (Ashby, 1990; Asher *et al.*, 2002; CIMMYT, 1988; Coe *et al.*, 2003; Dyke, 1974; Freeman, 2001; Gomez and Gomez, 1984; Mutsaers *et al.*, 1997; Patel *et al.*, 2004; Stroud, 1993; Tripp, 1982). Assessment of the validity of on-farm experiments thus appears incongruous with the method’s core feature and main strength: the ability to establish causal relationships through a repeatable procedure. Apparently, agronomists (reviewing on-farm studies) assess the validity of on-farm experimental findings using different criteria than the reproducibility of the followed scientific procedure. Agronomic textbooks are, however, largely silent on the alternative ways in which the validity of on-farm experiments is assessed.

Similarly, the external validity of experiments does not feature prominently in agronomic textbooks. First, the concept itself, appears – unlike in the social sciences (Leviton, 2015) – not well-known in agronomy; it does not feature in agronomic textbooks. Second, the extent to which

related topics such as the research population, the representativeness of sites and farmers, etc., are discussed varies greatly. Some textbooks give very little information regarding the contextual information that needs to be considered or may not even discuss the research population (e.g. Dyke, 1974; Patel *et al.*, 2004). Other textbooks address the issue more extensively (e.g. Mutsaers *et al.*, 1997; Stroud, 1993; Tripp, 1982). This greater concern for representativeness and the wider applicability of on-farm experiments may be related to the upsurge of FSR in the 1980s and 1990s.

When addressing the wider applicability of experimental research, textbooks written within the FSR tradition, typically put more emphasis on the social dimension. The research population is often described as a group or category of farmers with similar features (CIMMYT, 1988; Stroud, 1993; Tripp, 1982) and may be referred to as: the target group (Stroud, 1993), the target population (Mutsaers *et al.*, 1997), target area (Gomez and Gomez, 1984), recommendation domain (Tripp, 1982) and research domain (CIMMYT, 1988). Interestingly, none of the textbooks consider defining the size of the research population.

Although agronomic textbooks emphasise the need to select sites that are representative of the research population (e.g. Asher *et al.*, 2002; Dyke, 1974; Patel *et al.*, 2004), how to ensure representativeness is hardly discussed. For instance, Stroud (1993) suggests that '*On-farm trials are more meaningful if conducted with representative farms on representative sites*'. While textbooks may provide more guidance on the selection of trial-hosting farmers – for instance, by suggesting the inclusion of '*innovative farmers*', '*farmers with good communication abilities*' or '*experienced farmers*' (Ashby, 1990) – the consequences of such selection procedures for the experimental findings and their wider applicability are hardly reflected upon (Ashby, 1990; Freeman, 2001). Scant attention is generally given to the importance of transparency concerning the selection procedures and their justification. Where the problem of selection bias is acknowledged (e.g. Freeman, 2001; Tripp, 1982), research strategies to deal with such bias are not usually elaborated upon. For instance, one way to check for selection biases is to compare the selected trial-hosting farmers with the farmers in the wider research population, as suggested by Coe *et al.* (2003).

Limited guidance in agronomic textbooks and the half-hearted attempts in on-farm studies to define the research population, the lack of transparency about farmer selection procedures and reflection on the consequence of possible selection bias, the use of standardized 'farmer practice' treatments, etc., as discussed in this paper, suggest that agronomists publishing on on-farm experiments pay insufficient attention to the wider relevance of their work. Apparently, an explanation of the broader applicability is not a major concern in the assessment of a study's value or justification of its implementation. This is not merely remarkable in view of the emphasis on research impact and 'impact at scale' among funders of development-oriented agronomic research in Africa. It also undermines the relevance and, potentially, the future of on-farm experimental studies. After all, why would one – continue to – invest in on-farm experimental research if results are highly localised and their wider applicability unknown?

Conclusion

Towards more systematic description and strategic use of on-farm experiments

This review of on-farm experimental studies in Africa revealed that published studies generally provide insufficient information to reproduce the trials described in them. Predominantly using yield as the dependent variable, the description of the experimental design captures best the nutrient management, weed management and crop information, but pays little attention to other yield-determining factors, such as field history, land preparation and management of pests, weeds and water. The procedure developed in this paper to assess on-farm studies can be used to more systematically describe relevant experimental conditions. This may assist assessments of the validity of on-farm experiments and to increase the reproducibility of such experiments. Many

academic journals now offer ‘supplementary materials’ sections to authors, in which details on these 7 categories of yield-determining factors could be included.

The wider applicability or external validity of on-farm experiments is often also poorly addressed, even in studies that use stratified or experimentation at scale approaches. Most studies fail to explicitly define and describe the research population and/or environment in which (they expect) the experimental finding to work. We propose that experimental studies should minimally address: (1) under what conditions the treatment effect occurs and (2) for whom the experimental findings are relevant. Agronomy journals could include these suggestions in their guidelines for reviewers. In addition, we advocate for more transparency about the selection criteria used for research site, farmer and field selection and a reflection on the possible implications of such selection for the experiment’s external validity. Especially studies using a ‘farmer practice’ treatment should critically reflect on what (diverse) farming realities this researcher-constructed treatment is representative of.

The poor reporting on experimental conditions and the wider applicability of experimental findings are perhaps reminiscent of an era in which most experimental work was conducted on research stations and had different aims. As agronomic textbooks provide little guidance on either reproducibility or external validity of on-farm experiments, better guidelines are needed on how to increase the reproducibility and wider applicability (external validity) of on-farm experimentation.

Supplementary material. To view supplementary material for this article, please visit <https://doi.org/10.1017/S0014479720000174>

Acknowledgments. Any opinions, findings, conclusion, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of CRP MAIZE, CRP WHEAT, or CIMMYT. We thank three anonymous referees for their comments which helped us to revise our analysis and conclusions.

Financial support. Ken Giller is grateful for a grant from the NWO-WOTRO Strategic Partnership NL-CGIAR. This work was partly funded by the CGIAR Research Programs MAIZE (www.maize.org) and WHEAT (www.wheat.org) coordinated by the International Maize and Wheat Improvement Center (CIMMYT) in Mexico.

References

- Abdalla E.A., Osman A.K., Maki M.A., Nur F.M., Ali S.B. and Aune J.B. (2015). The response of sorghum, groundnut, sesame, and cowpea to seed priming and fertilizer micro-dosing in South Kordofan state, Sudan. *Agronomy-Basel* 5(4), 476–490.
- Anderson G. (1974). Bean responses to fertilizers on Mt. Kilimanjaro in relation to soil and climatic conditions. *East African Agricultural and Forestry Journal* 39(3), 272–288.
- Andersson J.A., Krupnik T.J. and de Roo N. (2019). On-farm trials as ‘Infection Points’? A response to Wall et al. *Experimental Agriculture* 55(2), 195–199.
- Ashby J.A. (1990). *Evaluating Technology with Farmers: A Handbook*. Cali, Colombia: CIAT.
- Asher C., Grundon N. and Menzies N. (2002). *How to Unravel and Solve Soil Fertility Problems*. Canberra: The Australian Centre for International Agricultural Research (ACIAR).
- Baker M. (2016). Reproducibility crisis. *Nature* 533, 26.
- Baudron F., Tittone P., Corbeels M., Letourmy P. and Giller K.E. (2012). Comparative performance of conservation agriculture and current smallholder farming practices in semi-arid Zimbabwe. *Field Crops Research* 132, 117–128.
- Becker M. and Johnson D.E. (2001a). Cropping intensity effects on upland rice yield and sustainability in West Africa. *Nutrient Cycling in Agroecosystems* 59(2), 107–117.
- Becker M. and Johnson D.E. (2001b). Improved water control and crop management effects on lowland rice productivity in West Africa. *Nutrient Cycling in Agroecosystems* 59(2), 119–127.
- Bielders C.L. and Gérard B. (2015). Millet response to microdose fertilization in south–western Niger: Effect of antecedent fertility management and environmental factors. *Field Crops Research* 171, 165–175.
- Brocke K.V., Gilles T., Weltzien E., Barro-Kondombo C.P., Goze E. and Chanterau J. (2010). Participatory variety development for sorghum in Burkina Faso: Farmers’ selection and farmers’ criteria. *Field Crops Research* 119(1), 183–194.
- Bucheyeki T.L., Shenkalwa M.E., Mapunda X.T. and Matata W.L. (2010). The groundnut client oriented research in Tabora, Tanzania. *African Journal of Agricultural Research* 5(5), 356–362.

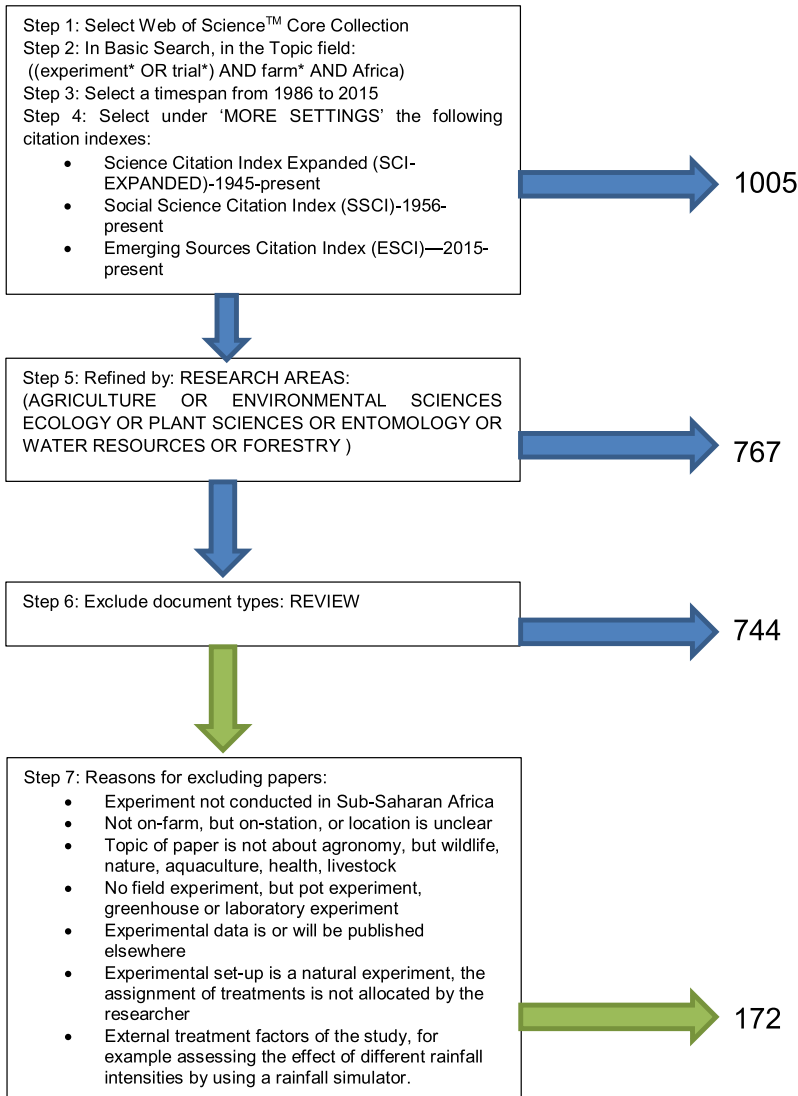
- Byerlee D., Harrington L. and Winkelmann D.L. (1982). Farming systems research: Issues in research strategy and technology design. *American Journal of Agricultural Economics* **64**(5), 897–904.
- Campbell D.T. and Stanley J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.
- Chambers R. and Jiggins J. (1987). Agricultural research for resource-poor farmers Part I: Transfer-of-technology and farming systems research. *Agricultural Administration and Extension* **27**(1), 35–52.
- CIMMYT (1988). *From Agronomic Data to Farmer Recommendations: An Economics Training Manual*. Completely revised edition. Mexico, DF: CIMMYT.
- Coe R., Franzel S., Beniast J. and Barahona C. (2003). *Designing Participatory On-Farm Experiments. A Resource Pack for Training*. Nairobi: World Agroforestry Centre.
- Coe R., Sinclair F. and Barrios E. (2014). Scaling up agroforestry requires research ‘in’ rather than ‘for’ development. *Current Opinion in Environmental Sustainability* **6**, 73–77.
- Collinson M.P. (2000). *A History of Farming Systems Research*. Wallingford: CABI.
- de Roo N., Andersson J.A. and Krupnik T.J. (2019). On-farm trials for development impact? The organisation of research and the scaling of agricultural technologies. *Experimental Agriculture* **55**(2), 163–184.
- Dyke G.V. (1974). *Comparative Experiments with Field Crops*. London, UK: Butterworths.
- Ebanyat P., de Ridder N., de Jager A., Delve R.J., Bekunda M.A. and Giller K.E. (2010). Impacts of heterogeneity in soil fertility on legume-finger millet productivity, farmers’ targeting and economic benefits. *Nutrient Cycling in Agroecosystems* **87**(2), 209–231.
- Falconnier G.N., Descheemaeker K., Van Mourik T.A. and Giller K.E. (2016). Unravelling the causes of variability in crop yields and treatment responses for better tailoring of options for sustainable intensification in southern Mali. *Field Crops Research* **187**, 113–126.
- Fanadzo M., Chiduzo C. and Mkeni P.N.S. (2010). Comparative performance of direct seeding and transplanting green maize under farmer management in small scale irrigation: A case study of Zanyokwe, Eastern Cape, South Africa. *African Journal of Agricultural Research* **5**(7), 524–531.
- Fox P. and Rockström J. (2000). Water-harvesting for supplementary irrigation of cereal crops to overcome intra-seasonal dry-spells in the Sahel. *Physics and Chemistry of the Earth Part B-Hydrology Oceans and Atmosphere* **25**(3), 289–296.
- Franke A.C., Berkhout E.D., Iwuafor E.N.O., Nziguheba G., Dercon G., Vandeplas I. and Diels J. (2010). Does crop-livestock integration lead to improved crop production in the savanna of West Africa? *Experimental Agriculture* **46**(4), 439–455.
- Freeman H.A. (2001). *Comparison of Farmer-Participatory Research Methodologies: Case Studies in Malawi and Zimbabwe*. Nairobi, Kenya: Socioeconomics and Policy Program, International Crops Research Institute for the Semi-Arid Tropics.
- Gildemacher P.R., Schulte-Geldermann E., Borus D., Demo P., Kinyae P., Mundia P. and Struik P.C. (2011). Seed potato quality improvement through positive selection by smallholder farmers in Kenya. *Potato Research* **54**(3), 253–266.
- Giller K.E., Tittone P., Rufino M.C., van Wijk M.T., Zingore S., Mapfumo P., Adjei-Nsiah S., Herrero M., Chikowo R., Corbeels M., Rowe E.C., Bajjukya F., Mwijage A., Smith J., Yeboah E., van der Burg W.J., Sanogo O.M., Misiko M., de Ridder N., Karanja S., Kaizzi C., K’ungu J., Mwale M., Nwaga D., Pacini C. and Vanlauwe B. (2011). Communicating complexity: integrated assessment of trade-offs concerning soil fertility management within African farming systems to support innovation and development. *Agricultural Systems* **104**(2), 191–203.
- Giller K.E., Franke A.C., Abaidoo R., Bajjukya F.P., Bala A., Boahen S., Dashiell K., Katengwa S., Sanginga J., Sanginga N., Simmons A., Turner A., Woomer P.L., Wolf J.D. and Vanlauwe B. (2013). N2Africa: Putting nitrogen fixation to work for smallholder farmers in Africa. In Vanlauwe B., van Asten P. and Blomme G. (eds), *Agro-ecological Intensification of Agricultural Systems in the African Highlands*. London: Routledge, pp. 156–174.
- Gomez K.A. and Gomez A.A. (1984). *Statistical Procedures for Agricultural Research*. New York: John Wiley & Sons.
- John L.K., Loewenstein G. and Prelec D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* **23**(5), 524–532.
- Kaizzi K.C., Byalebeka J., Semalulu O., Alou I., Zimwanguyizza W., Nansamba A., Musinguzi P., Ebanyat P., Hyuha T. and Wortmann C.S. (2012). Maize response to fertilizer and nitrogen use efficiency in Uganda. *Agronomy Journal* **104**(1), 73–82.
- Kearney S.F., Steven J; Salomon Abraham; Six Johan; Scow Kate M (2012). Forty percent revenue increase by combining organic and mineral nutrient amendments in Ugandan smallholder market vegetable production. *Agronomy for Sustainable Development* **32**(4), 831–839.
- Khan Z.R., Midega C.A.O., Hassanali A., Pickett J.A., Wadhams L.J. and Wanjoya A. (2006). Management of witchweed, *Striga hermonthica*, and stemborers in sorghum, *Sorghum bicolor*, through intercropping with greenleaf desmodium, *Desmodium intortum*. *International Journal of Pest Management* **52**(4), 297–302.
- Krupnik T.J., Shennan C., Settle W.H., Demont M., Ndiaye A.B. and Rodenburg J. (2012). Improving irrigated rice production in the Senegal River Valley through experiential learning and innovation. *Agricultural Systems* **109**, 101–112.
- Lamers J.P.A., Bruentrup M. and Buerkert A. (2015). Financial performance of fertilization strategies for sustainable soil fertility management in Sudano-Sahelian West Africa. 2: Profitability of long-term capital investments in rockphosphate. *Nutrient Cycling in Agroecosystems* **102**(1), 149–165.

- Leviton L.C. (2015). External validity. In James D. Wright (eds), *International Encyclopedia of the Social & Behavioral Sciences*, 2nd Edn. Oxford: Elsevier, pp. 617–622.
- Maat H. (2011). The history and future of agricultural experiments. *Njas-Wageningen Journal of Life Sciences* 57(3–4), 187–195.
- Manu-Aduening J.A., Lamboll R.I., Mensah G.A., Lamptey J.N., Moses E., Dankyi A.A. and Gibson R.W. (2006). Development of superior cassava cultivars in Ghana by farmers and scientists: The process adopted, outcomes and contributions and changed roles of different stakeholders. *Euphytica* 150(1–2), 47–61.
- Misiko M., Tittone P., Ramisch J.J., Richards P., Giller K.E. (2008). Integrating new soybean varieties for soil fertility management in smallholder systems through participatory research: Lessons from western Kenya. *Agricultural Systems* 97(1–2), 1–12.
- Morse S., McNamara N. and Acholo M. (2009). Potential for clean yam miniset production by resource-poor farmers in the middle-belt of Nigeria. *Journal of Agricultural Science* 147, 589–600.
- Mucheru-Muna M., Mugendi D., Pypers P., Mugwe J., Kung'u J., Vanlauwe B. and Merckx R. (2014). Enhancing maize productivity and profitability using organic inputs and mineral fertilizer in central Kenya small-hold farms. *Experimental Agriculture* 50(2), 250–269.
- Mutsaers H., Weber G., Walker P. and Fisher N. (1997). *A Field Guide for On-Farm Experimentation*. Ibadan: IITA.
- Ncube B., Dimes J.P., Twomlow S.J., Mupangwa W. and Giller K.E. (2007). Raising the productivity of smallholder farms under semi-arid conditions by use of small doses of manure and nitrogen: A case of participatory research. *Nutrient Cycling in Agroecosystems* 77(1), 53–67.
- Ndjeunga J. and Bationo A. (2005). Stochastic dominance analysis of soil fertility restoration options on sandy Sahelian soils in southwest Niger. *Experimental Agriculture* 41(2), 227–244.
- Nyakudya I.W., Stroosnijder L. and Nyagumbo I. (2014). Infiltration and planting pits for improved water management and maize yield in semi-arid Zimbabwe. *Agricultural Water Management* 141, 30–46.
- Nyamangara J., Nyengerai K., Masvaya E.N., Tirivavi R., Mashingaidze N., Mupangwa W., Dimes J., Hove L. and Twomlow S. (2014). Effect of conservation agriculture on maize yield in the semi-arid areas of Zimbabwe. *Experimental Agriculture* 50(2), 159–177.
- Ojiem J.O., Franke A.C., Vanlauwe B., de Ridder N. and Giller K.E. (2014). Benefits of legume-maize rotations: Assessing the impact of diversity on the productivity of smallholders in Western Kenya. *Field Crops Research* 168, 75–85.
- Ojiem J.O., Vanlauwe B., de Ridder N. and Giller K.E. (2007). Niche-based assessment of contributions of legumes to the nitrogen economy of Western Kenya smallholder farms. *Plant and Soil* 292(1–2), 119–135.
- Pandey R.K., Maranville J.W. and Crawford T.W. (2001). Agriculture intensification and ecologically sustainable land use systems in Niger: Transition from traditional to technologically sound practices. *Journal of Sustainable Agriculture* 19(2), 5–24.
- Pashler H. and Wagenmakers E.J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science* 7(6), 528–530.
- Patel B., Muir-Leresche K., Coe R. and Hainsworth S. (2004). *The Green Book: a Guide to Effective Graduate Research in African Agriculture, Environment and Rural Development*. Kampala, Uganda: The African Crop Science Society.
- Pypers P., Bimponda W., Lodi-Lama J.-P., Lele B., Mulumba R., Kachaka C., Boeckx P., Merckx R. and Vanlauwe B. (2012). Combining mineral fertilizer and green manure for increased, profitable cassava production. *Agronomy Journal* 104(1), 178–187.
- Ratnadas A., Cisse B., Cisse S., Cisse T., Hamada M.A., Chantreau J. and Letourmy P. (2008). Combined on-farm effect of plot size and sorghum genotype on sorghum panicle-feeding bug infestation in Mali. *Euphytica* 159(1–2), 135–144.
- Richards P. (1985). *Indigenous Agricultural Revolution: Ecology and Food Production in West Africa*. Boulder, CO: Westview Press.
- Rockström J., Kaurnbutho P., Mwalley J., Nzabi A.W., Temesgen M., Mawenya L., Barron J., Mutua J. and Damgaard-Larsen S. (2009). Conservation farming strategies in East and Southern Africa: Yields and rain water productivity from on-farm action research. *Soil & Tillage Research* 103(1), 23–32.
- Ronner E., Franke A., Vanlauwe B., Dianda M., Edeh E., Ukem B., Bala A., Van Heerwaarden J. and Giller K.E. (2016). Understanding variability in soybean yield and response to P-fertilizer and rhizobium inoculants on farmers' fields in northern Nigeria. *Field Crops Research* 186, 133–145.
- Roothaert R., Franzel S. and Kiura M. (2003). On-farm evaluation of fodder trees and shrubs preferred by farmers in central Kenya. *Experimental Agriculture* 39(4), 423–440.
- Sanou H., Sidibé D., Korbo A. and Teldehaimanot Z. (2014). Rootstock Propagation Methods Affect the Growth and Productivity of Three Improved Cultivars of Ber in Mali, West Africa. *Horttechnology* 24(4), 418–423.
- Scaife M. (1968). Maize fertilizer experiments in Western Tanzania. *The Journal of Agricultural Science* 70(2), 209–222.
- Snapp S.S., Rohrbach D.D., Simtowe F. and Freeman H.A. (2002). Sustainable soil management options for Malawi: Can smallholder farmers grow more legumes? *Agriculture Ecosystems & Environment* 91(1–3), 159–174.
- Stroud A. (1993). *Conducting On-Farm Experiments*. Cali, Colombia: CIAT.

- Thierfelder C., Matemba-Mutasa R. and Rusinamhodzi L.** (2015). Yield response of maize (*Zea mays* L.) to conservation agriculture cropping system in Southern Africa. *Soil & Tillage Research* **146**, 230–242.
- Tittonell P. and Giller K.E.** (2013). When yield gaps are poverty traps: The paradigm of ecological intensification in African smallholder agriculture. *Field Crops Research* **143**, 76–90.
- Tripp R.** (1982). *Data Collection, Site Selection and Farmer Participation in On-Farm Experimentation*. Mexico: CIMMYT.
- Trutmann P. and Graf W.** (1993). The impact of pathogens and arthropod pests on common bean production in Rwanda. *International Journal of Pest Management* **39**(3), 328–333.
- Tulema B., Aune J.B. and Breland T.A.** (2007). Availability of organic nutrient sources and their effects on yield and nutrient recovery of tef *Eragrostis tef* (Zucc.) Trotter and on soil properties. *Journal of Plant Nutrition and Soil Science-Zeitschrift Fur Pflanzenernahrung Und Bodenkunde* **170**(4), 543–550.
- Twomlow S., Rohrbach D., Dimes J., Rusike J., Mupangwa W., Ncube B., Hove L., Moyo M., Mashingaidze N. and Mahposa P.** (2010). Micro-dosing as a pathway to Africa's Green Revolution: Evidence from broad-scale on-farm trials. *Nutrient Cycling in Agroecosystems* **88**(1), 3–15.
- van Ittersum M.K. and Rabbinge R.** (1997). Concepts in production ecology for analysis and quantification of agricultural input-output combinations. *Field Crops Research* **52**(3), 197–208.
- Vanlauwe B., Coe R. and Giller K.E.** (2019). Beyond averages: New approaches to understand heterogeneity and risk of technology success or failure in smallholder farming. *Experimental Agriculture* **55**, 84–106.
- Vanlauwe B., Tittonell P. and Mukalama J.** (2006). Within-farm soil fertility gradients affect response of maize to fertiliser application in western Kenya. *Nutrient Cycling in Agroecosystems* **76**(2–3), 171–182.
- Waddington S.R., Karigwindi J. and Chifamba J.** (2007). The sustainability of a groundnut plus maize rotation over 12 years on smallholder farms in the sub-humid zone of Zimbabwe. *African Journal of Agricultural Research* **2**(8), 342–348.
- Worou O.N., Gaiser T., Saito K., Goldbach H. and Ewert F.** (2013). Spatial and temporal variation in yield of rainfed lowland rice in inland valley as affected by fertilizer application and bunding in North-West Benin. *Agricultural Water Management* **126**, 119–124.
- Yamoah C.F., Bationo A., Shapiro B. and Koala S.** (2011). Use of rainfall indices to analyze the effects of phosphate rocks on millet in the Sahel. *African Journal of Agricultural Research* **6**(3), 586–593.
- Zingore S., Murwira H., Delve R. and Giller K.E.** (2007). Soil type, historical management and current resource allocation: Three dimensions regulating variability of maize yields and nutrient use efficiencies on African smallholder farms. *Field Crop Research* **101**, 296–305.

Appendix I. Search in Web of Science (v.5.22.3)

Search was conducted on October 28, 2016.



Remark:

Selecting 'All databases' and refine to 'Web of Science™ Core Collection' gives 5 times as many records as selecting the 'Web of Science™ Core Collection' right at the start. The reason for this difference is as follows: an All Databases search includes additional keyword- and classification-based indexing fields from other databases, which are taken into account to retrieve articles. For example, an article can appear in both the Web of Science™ Core Collection and in other databases. Although the search criteria are not met in the Web of Science™ Core Collection version of the record, other databases may have a different version of the record, which matches with the search criteria. Thus this article will show up in an All databases search, refined to the Web of Science™ Core Collection, but not when conducting a search in the Web of Science™ Core Collection.

Cite this article: Kool H, Andersson JA, and Giller KE (2020). Reproducibility and external validity of on-farm experimental research in Africa. *Experimental Agriculture* 56, 587–607. <https://doi.org/10.1017/S0014479720000174>