Environ Dev Sustain
DOI 10.1007/s10668-016-9803-x

When experts disagree: the need to rethink indicator selection for assessing sustainability of agriculture

Evelien M. de Olde[1,2] · Henrik Moller[3] · Fleur Marchand[4,5] · Richard W. McDowell[6,7] · Catriona J. MacLeod[8] · Marion Sautier[3,9] · Stephan Halloy[10,11] · Andrew Barber[12] · Jayson Benge[12] · Christian Bockstaller[13,14] · Eddie A. M. Bokkers[2] · Imke J. M. de Boer[2] · Katharine A. Legun[15] · Isabelle Le Quellec[12] · Charles Merfield[16] · Frank W. Oudshoorn[1,17] · John Reid[18] · Christian Schader[19] · Erika Szymanski[20] · Claus A. G. Sørensen[1] · Jay Whitehead[21] · Jon Manhire[12]

Received: 4 March 2016 / Accepted: 29 April 2016

Abstract Sustainability indicators are well recognized for their potential to assess and monitor sustainable development of agricultural systems. A large number of indicators are proposed in various sustainability assessment frameworks, which raises concerns regarding the validity of approaches, usefulness and trust in such frameworks. Selecting indicators requires transparent and well-defined procedures to ensure the relevance and validity of sustainability assessments. The objective of this study, therefore, was to determine whether

✉ Evelien M. de Olde
evol@eng.au.dk

[1] Department of Engineering, Aarhus University, Inge Lehmanns Gade 10, 8000 Aarhus, Denmark

[2] Animal Production Systems Group, Wageningen University, Wageningen, The Netherlands

[3] Centre for Sustainability: Agriculture, Food, Energy, Environment, University of Otago, Dunedin, New Zealand

[4] Social Sciences Unit, Institute for Agricultural and Fisheries Research (ILVO), Merelbeke, Belgium

[5] Ecosystem Management Research Group and IMDO, University of Antwerp, Wilrijk, Belgium

[6] Invermay Agricultural Centre, AgResearch, Mosgiel, New Zealand

[7] Agriculture and Life Sciences, Lincoln University, Lincoln, New Zealand

[8] Landcare Research, Dunedin, New Zealand

[9] INRA, UMR 1248 AGIR, Castanet-Tolosan, France

[10] Universidad Nacional de Chilecito, La Rioja, Argentina

[11] Ministry for Primary Industries, Wellington, New Zealand

[12] The Agribusiness Group, Lincoln University, Lincoln, New Zealand

Published online: 11 May 2016

experts agree on which criteria are most important in the selection of indicators and indicator sets for robust sustainability assessments. Two groups of experts (Temperate Agriculture Research Network and New Zealand Sustainability Dashboard) were asked to rank the relative importance of eleven criteria for selecting individual indicators and of nine criteria for balancing a collective set of indicators. Both ranking surveys reveal a startling lack of consensus amongst experts about how best to measure agricultural sustainability and call for a radical rethink about how complementary approaches to sustainability assessments are used alongside each other to ensure a plurality of views and maximum collaboration and trust amongst stakeholders. To improve the transparency, relevance and robustness of sustainable assessments, the context of the sustainability assessment, including prioritizations of selection criteria for indicator selection, must be accounted for. A collaborative design process will enhance the acceptance of diverse values and prioritizations embedded in sustainability assessments. The process by which indicators and sustainability frameworks are established may be a much more important determinant of their success than the final shape of the assessment tools. Such an emphasis on process would make assessments more transparent, transformative and enduring.

Keywords Indicator selection · Multi-criteria assessment · Ranking · Sustainability assessment · Temperate agriculture

1 Introduction

Current concerns regarding global food security, climate change, animal welfare, biodiversity and availability of natural resources emphasize the need for sustainable development of agriculture (OECD 2001; Steinfeld et al. 2006; IAASTD 2009; Pretty et al. 2010). Although interpretation of the concept of sustainable development (i.e. further referred to as sustainability) varies widely, a consensus exists on the need to use relevant sustainability indicators to assess change (Hansen 1996; Bell and Morse 2008; Bockstaller et al. 2015). Sustainability indicators measure the current status of a system to identify trends, forewarning the breach of critical thresholds and monitoring the success of interventions to build sustainability.

[13] INRA, UMR 1121 Agronomie et Environnement, INRA-Université de Lorraine, BP20507, Colmar Cedex, France

[14] UMR 1121, Agronomie et Environnement, Université de Lorraine, BP20507, Colmar Cedex, France

[15] Department of Sociology, Gender and Social Work, University of Otago, Dunedin, New Zealand

[16] The BHU Future Farming Centre, Lincoln, New Zealand

[17] SEGES, Aarhus N, Denmark

[18] Ngai Tahu Research Centre, University of Canterbury, Christchurch, New Zealand

[19] Research Institute of Organic Agriculture (FiBL), Frick, Switzerland

[20] Centre for Science Communication, University of Otago, Dunedin, New Zealand

[21] Agribusiness and Economics Research Unit, Lincoln University, Lincoln, New Zealand

A wide range of indicator-based tools for sustainability assessment have been developed to assess the sustainability performance of agricultural systems (FAO 2013; Keulen et al. 2005; de Olde et al. 2016; Marchand et al. 2014). These tools vary in their assessment objective, spatial and temporal scope and level of stakeholder involvement (Binder et al. 2010; Schader et al. 2014). Consensus on which sustainability indicators to include is lacking and contributes to a wide diversity of approaches (Bockstaller et al. 2009; Bell and Morse 2008; Parris and Kates 2003). This multiplicity can add cost, impair the ability to focus on the most salient sustainability indicators and raise concerns regarding the validity of approaches, usefulness and trust in the concept of sustainability (Hansen 1996; Bockstaller et al. 2009; Schader et al. 2014). As a solution, several have raised the importance of transparent and well-defined procedures and criteria for selecting individual indicators and balancing indicator sets to develop relevant, trusted, comprehensible and meaningful sustainability assessments (Dale and Beyeler 2001; Bockstaller et al. 2009; Niemeijer and de Groot 2008; Lebacq et al. 2013). This paper, therefore, focuses on the general criteria for indicator selection as an overarching issue in the design of sustainability assessments and as a key step before defining individual indicators and assessment methods (Reed et al. 2006; Dale and Beyeler 2001).

The selection of sustainability indicators is made using a list of criteria (Dale and Beyeler 2001). Criteria for the selection of individual sustainability indicators discussed in the scientific literature commonly cover relevance, validity, measurability, sensitivity and comprehensibility by stakeholders and decision-makers (Dale and Beyeler 2001; Lebacq et al. 2013). Together, the collective set of indicators should comprehensively represent the agricultural system (Niemeijer and de Groot 2008; Binder et al. 2010; Marchand et al. 2014). The definition, selection and prioritization of selection criteria used to select indicators vary widely between sustainability assessment tools (Niemeijer and de Groot 2008; Bell and Morse 2008; Reed et al. 2006). Describing criteria used for selecting indicators is, therefore, important for the transparency and reliability of sustainability assessments (Dale and Beyeler 2001; Niemeijer and de Groot 2008).

This paper discusses the results of a ranking survey amongst experts on sustainability assessment of agricultural systems, regarding the relative importance of criteria to select individual indicators and balance a collective set of indicators. The objective was to determine whether these experts agree on which criteria are most important, and if not, discuss the implications for building reliable sustainability assessments in the future.

2 Methods

To get insight into criteria, principles and processes to build a reliable sustainability assessment, we started with an overview of eleven criteria for individual indicator selection (Table 1) and nine criteria for balancing the collective set of indicators (Table 2) to assess the sustainability performance of agricultural systems. These criteria were judged to be the most important (based on their emphasis in the sustainability monitoring literature) by Moller and MacLeod (2013) from their review of international and New Zealand sustainability assessment initiatives in agriculture and ecology (Lee et al. 2005; OECD 2001; Sommerville et al. 2011; Herzog et al. 2012; Jones et al. 2012). To make the ranking

process tractable, we selected and summarized the broad spectrum of criteria used to define indicators. Many criteria are listed in the sustainability literature, and elements of those we chose are grouped and framed in different ways (Moller and MacLeod 2013). Accordingly, our survey should not be perceived as a definitive list of all the potential criteria to be considered. Tables 1 and 2 present the criteria and descriptions exactly as described to the participants in the survey. The two groups were:

1. Invited members of Pilot Activity 1 (Resilient Agriculture Production Systems) in the recently launched international Temperate Agriculture Research Network (TempAg). TempAg participants were all experts in sustainability assessment of agricultural systems and consisted of researchers or agricultural policy analysts. Participants were selected based on their expertise in temperate agriculture and sustainability assessments and representation of different geographical areas and disciplines (i.e. economy, ecology, policy and social science). Eighteen respondents ranked the criteria according to the goals of the TempAg network to support sustainable agriculture over a wide range of production sectors, temperate biomes and sociopolitical systems. Panel members were from Argentina, Australia, Austria, Belgium, Canada, Denmark, France, Italy, Japan, the Netherlands, New Zealand, Sweden, Switzerland and the USA.
2. Twenty members of the New Zealand Sustainability Dashboard (NZSD) research team. This team, a coalition of researchers and consultants, was asked to complete the same ranking surveys while focusing on the specific needs of the NZSD. The dashboard is a package of tools that deploys an industry-led approach to measuring and reporting sustainability at the farm level in New Zealand. It uses a participatory approach in which the involvement of stakeholders is contractual. It allows farmers to log self-assessed sustainability measures into an online database (Merfield et al. 2015). The NZSD panel, therefore, assessed the criteria in terms of a narrower defined context than that of the newly formed and as yet not fully defined TempAg agenda and team. The panel included agronomists, farm advisors, ecologists, rural sociologists and economists. Half of the NZSD team are researchers based in universities, and half are professional agricultural consultants.

In view of the broad diversity of TempAg and NZSD participants, we anticipated our results to be broadly applicable across agricultural sustainability concerns. In total, 38 participants each spent approximately 25 min to complete the two surveys. The ranking survey was carried out in April and May 2015, using an online decision-making software package (1000Minds). Agreement in ranking of participants was tested using the non-parametric Kendall's W (coefficient of concordance) in which 1 indicates perfect agreement between participants, and 0 indicates a complete lack of agreement (Siegel 1956; Kendall and Smith 1939; Gibbons and Chakraborti 2011). To test differences in the rank scores between researchers and consultants of the NZSD team, an unpaired, two-sample Wilcoxon test was run in R.

Table 1 Possible criteria for selecting individual agricultural sustainability indicators, after Moller and MacLeod (2013)

Criterion	Description
Sustainability relevance	Indicators should measure key properties of environment, economy, society or governance that affect sustainability (e.g. state, pressure, response, use or capability)
Clearly defined and standardized	Indicators must be based on clearly defined, verifiable and scientifically acceptable data collected using standarzised methods so that they can be reliably repeated and compared against each other
Easily communicated and understood	Easily communicated and understood
Broad acceptance	The strength of an indicator depends on its broad acceptance by major stakeholders (e.g. growers, policy-makers, scientists, customers)
Affordable measurement	Affordable measurement increases participation and regularity of monitoring or broadens the scope of what can be measured for overall sustainability assessment
Performance rather than practice based	It is better to measure actual performance and outcomes rather than just practices that are expected to promote sustainability and resilience
Sensitivity	Indicators should be sensitive (change immediately and a lot if agricultural systems status changes). This helps detect trends or breaches of thresholds within the time frames and on the scales that are relevant to the management decisions, and before it is too late to correct any problems
Quantification	Indicators should be fully quantified whenever practicable. Counts and continuous variables (interval and ratio scales) are more favoured than ranks (ordinal scales) or 'yes/no' scores (binary); any form of quantification is preferable to a fully qualitative assessment
Specificity for interpretability	Indicators should be affected only by a few key drivers (risks, opportunities, causes) of sustainability rather than being affected by many things (local context, multiple stressors, etc.) in order for any change in the indicator to be interpretable for sustainability
High precision and statistical power	Indicators must have sufficient precision and accuracy and sufficiently low natural variance for monitoring to detect trends and probability that some limit or threshold has been breached
Capacity to upscale	Indicators should be designed and measured in a way that allows their aggregation at multiple spatial and temporal scales for different purposes

Table 2 Possible criteria for balancing the collective set of indicators for agricultural sustainability assessment, after Moller and MacLeod (2013)

Criterion	Description
Participatory co-development	Indicator sets and frameworks that are co-designed by key stakeholders are more likely to be relevant, trusted, practical, heeded and used for learning
Wide scope and integration	The framework and indicator sets must cover and cross-link multiple dimensions of sustainability and values encompassing environment, economics, social and governance dimensions
Linked to targets/ thresholds	Indicators should be linked to realizable, action-oriented, measurable and time-delimited targets or critical thresholds of risk, performance or best professional practice
Transparency and accessibility	Datasets that are accessible to all stakeholders (including the public) and explain assumptions, uncertainty and sources are more likely to be trusted and used

Table 2 continued

Criterion	Description
Policy relevant and meaningful	Indicators should send a clear message and provide information at an appropriate level for policy and management decision-making by assessing changes in the status of and risks to agricultural sustainability
Just enough indicators	The fewer the indicators, the better, provided the critical determinants of sustainability have been covered. Having just enough indicators will result in more participation, improved accuracy in reporting and clearer communication of the overall picture to farmers, policymakers and the public
Mix of generalized and specific	Indicator sets must include enough general indicators to allow cross-comparison between agricultural sectors, regions, countries and diverse social-ecological systems. However, some highly specific and locally grounded indicators must be included to guide fine-grained management adjustments that are especially relevant to one sector or region/country
Balance of current and future	Monitoring is part of risk management, so it must inform current options and drivers while preparing actors for future turbulence (shocks and drivers). At least some of the indicators and measurements should monitor potential new threats and opportunities just over the horizon
Explanatory and context info	Management guidance is more focused, effective and reliable and benchmarking more fair if additional information is gathered to identify covariates and additional information to determine why the indicators change

3 Results

Ranges in scores demonstrated a wide variation in perceived importance of selection criteria (Fig. 1). 'Sustainability relevance', the ability to measure environmental, economic, social and governance performance, was perceived, on average, the most important selection criterion for sustainability indicators. Criteria related to the 'acceptance', 'standardization' and 'communication' of the indicator were also ranked highly. Kendall's W was 0.31 for this ranking survey, a reflection of the weak consensus amongst the experts about how to best select individual indicators. A higher level of agreement was observed amongst NZSD participants ($W = 0.52$) than TempAg participants ($W = 0.23$).

Whereas the criteria 'specificity for interpretability' and 'sensitivity' received a substantially higher median rank by TempAg participants, the criteria 'affordable measurement' and 'easily communicated' were considered more important by NZSD (Fig. 1).

In the second ranking survey, on balancing a collective set of indicators, 'participatory co-development' achieving a 'wide scope and integration' and establishing 'links to targets and critical thresholds' were considered as the three most important criteria (Fig. 2).

Similar to the first survey, the results demonstrated a wide variation in the perceived importance of selection criteria for indicator sets. Each criterion received the full range of possible ranks (1–9) across the participants, indicating a lack of agreement amongst participants (Kendall's $W = 0.09$, with W-values for TempAg and NZSD of 0.10, and 0.15, respectively). As a result, median ranks of criteria remain between rank 3 and 6. This relatively flat trajectory of the median ranks illustrates the absence of consensus on priorities in balancing indicator sets (Fig. 2).

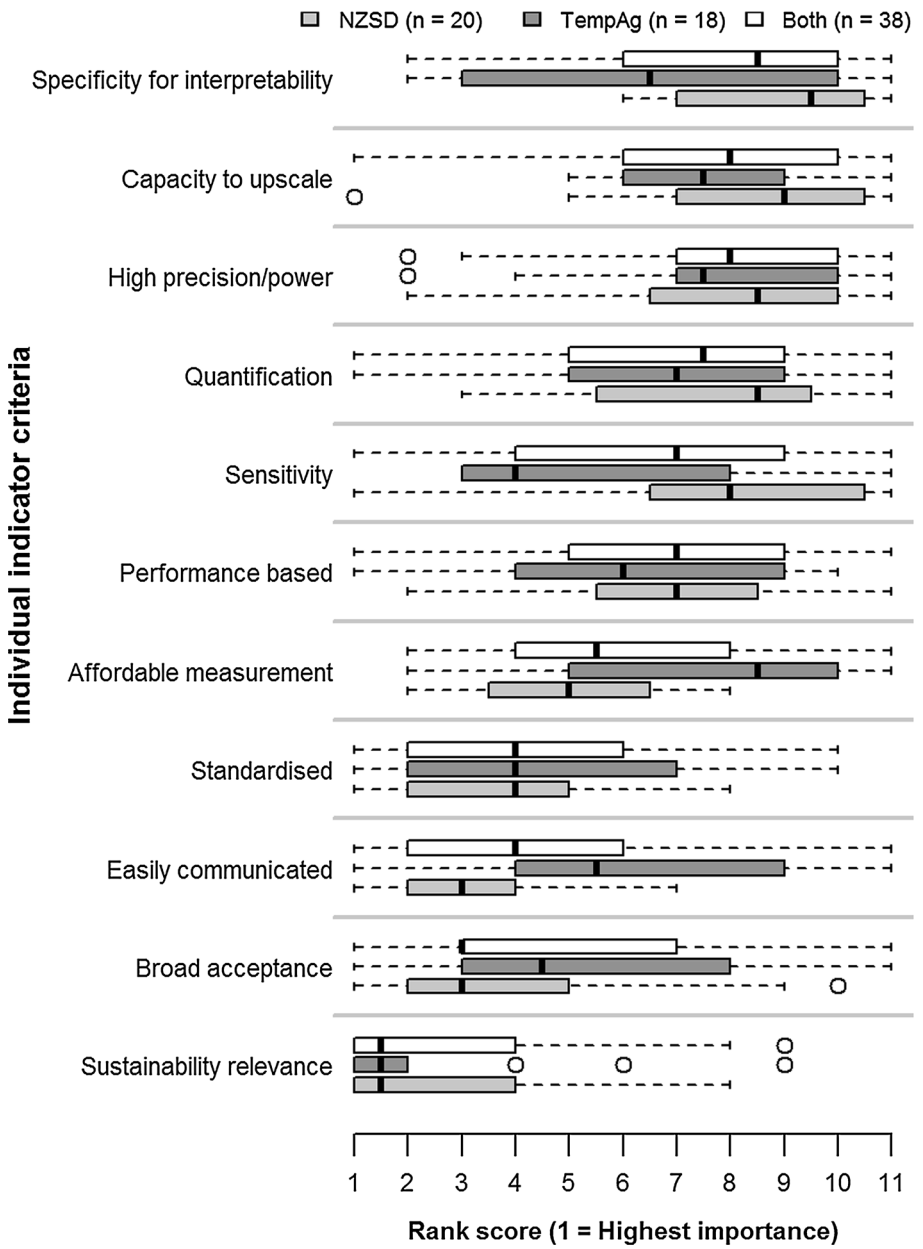

Fig. 1 Criteria for choosing agricultural sustainability indicators. The highest importance is indicated by rank 1 and the lowest by rank 11. Criteria are ranked in increasing order of importance measured by medians based on both surveys. Boxes contain the 25th and 75th percentiles, and the line within the *box* is the median. Whiskers extend to the most extreme data point (which is no more than 1.5 times the interquartile range from the *box*), and *outlier points* show the minimum and maximum values

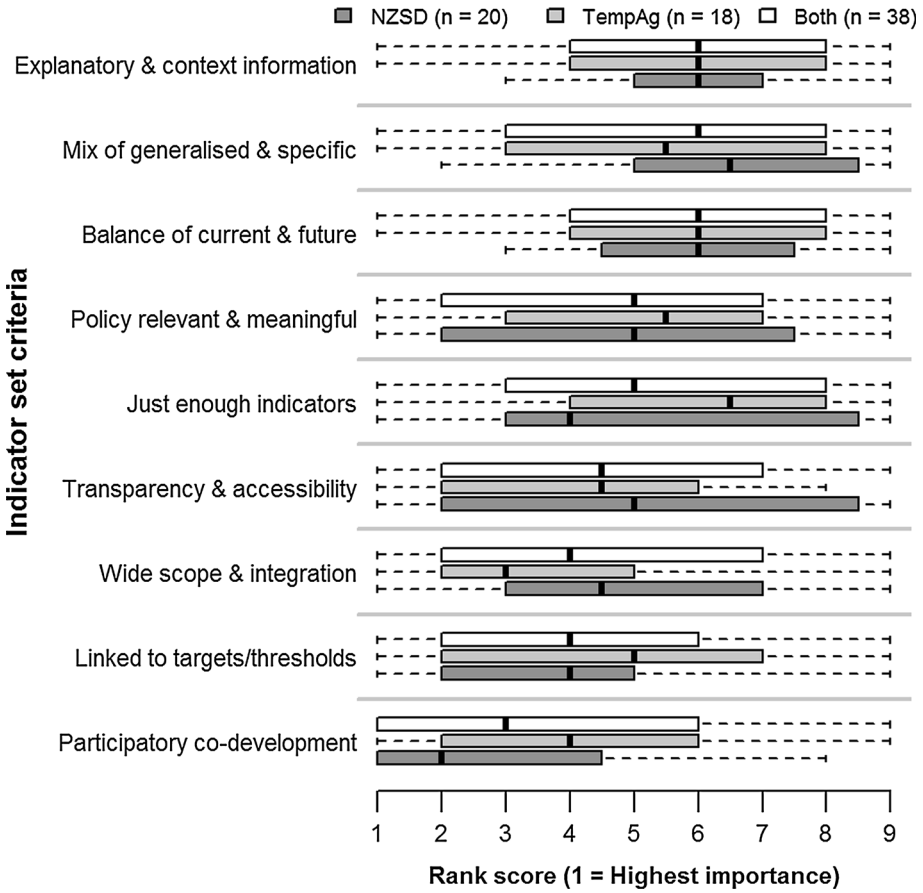

Fig. 2 Criteria for balancing indicator sets for agricultural sustainability assessment. Criteria are ranked in increasing order of importance measured by medians based on both surveys. *Boxes* contain the 25th and 75th percentiles, and the *line* within the box is the median. Whiskers extend to the most extreme data point (which is no more than 1.5 times the interquartile range from the *box*), and outlier points show the minimum and maximum values

TempAg and NZSD participants ranked criteria for balancing indicators sets quite differently (Fig. 2). NZSD participants tended to prioritize participatory co-development and keeping the number of indicators to a minimum, whereas TempAg participants emphasized the need for a wide scope and integration of indicator sets to cover multiple dimensions of sustainability.

Within the NZSD team, median rank scores for 'sensitivity' and, to a lesser extent, 'sustainability relevance' were higher for the researchers than for the consultants (Wilcoxon rank test; sensitivity: $W = 66$, $P = 0.002$; relevance: $W = 69.5$, $P = 0.046$). No other differences emerged between researchers and consultants in individual indicator selection criteria, nor in any of the criteria for balancing indicator sets.

4 Discussion

4.1 Lack of consensus, even amongst experts

The most remarkable feature of both ranking surveys is the lack of consensus amongst participants about what matters most in indicator selection criteria. This is shown by the low values of Kendall's *W*, the wide range of scores and the relatively flat trajectory of the median ranks, especially for the survey on indicator sets. A possible explanation could be that experts were targeting different types of indicators for different social and economic contexts, farming systems (e.g. confined vs. pasture-based livestock systems) and user groups (e.g. farmers or policy-makers). The first survey focusing on individual indicator selection had a greater level of agreement (based on the Kendall's *W* value) amongst NZSD participants compared to TempAg participants. Within the TempAg participants, researchers represented a wider variety of disciplines, from more diverse contexts. Findings for the NZSD participants, however, show that even in a team with a shared goal, the rankings vary strongly.

Differences in TempAg and NZSD prioritizations could be explained by their distinctive agendas. The high dependence of NZSD on farmers and industrial partners along with their commercial focus and participatory approach could explain their higher ranking of 'affordability' and 'easily communicated' criteria. By contrast, TempAg's goal to provide highly robust and technically derived indicators for comparison of agricultural performance between temperate countries in the OECD could explain their higher ranking of 'specificity' and 'sensitivity'.

Other factors that could have contributed to the apparent lack of consensus include the limitation of the ranking process to allow equal importance of criteria. Many of the participants may have judged some criteria to be about equally important, and the resulting ranking therefore forced artificial and inconsistent ordering of priorities. Other potential reasons for the lack of consensus include use of relatively broad descriptions for selection criteria and the diversity of nationalities and disciplinary backgrounds of the participants involved. The latter may have influenced the interpretation and understanding of the criteria (Lupia 2013).

As experts operating with different training and access to different knowledge sources, participants in the survey would be likely to weigh the importance of criteria differently (Dovers 2005). As a narrow focus may allow for greater vision in some areas, it may preclude us from fully seeing the importance of other indicator criteria. From this perspective, the divergence in the evaluation could be seen as an asset, as a range of expertise can generate a more rigorous exploration of indicators and sounder assessment development as multiple types of expertise are brought together. These differences in knowledge systems preface the call for interdisciplinarity and transdisciplinarity in sustainability research (Kates et al. 2001; Komiyama and Takeuchi 2006; Ostrom 2009; Popa et al. 2015). Not only can integration of diverse knowledge domains improve our understanding of sustainability issues, but it also can aid governance of those issues and the enhancement of democratic processes (Bäckstrand 2004). The persistent variation in prioritization of criteria used to define indicators can attest to the maintenance of specialized knowledge, even in interdisciplinary environments, and yet it presents challenges for implementing actionable sustainability programs.

Divergent views may reflect differences in world views, for example reductionist versus more holistic or system-oriented approaches to understanding a complex social-ecological activity like agriculture. A lack of consensus is perhaps expected when asking experts what

is more important to monitor within a complex and interconnected set of processes. Therefore, we emphasize the need to include a plurality of world views in a flexible framework for the selection of indicators. Participatory approaches which incorporate diverse views in the form of broad patterns, co-development of scenarios and use of adaptive planning approaches are all useful tools to build flexibility and inclusivity (Seimon et al. 2009, 2012; Yager et al. 2009).

4.2 Selecting indicators: the importance of context, plurality and flexibility

In our study, we have defined criteria for the selection of individual indicators and indicator sets based on a review of criteria used in agriculture and ecology (Moller and MacLeod 2013). The selection criteria are, however, generic and could be applied in other disciplines. We expect, however, that any complex adaptive system that demands management of human society, ecology and biology, will be confronted with similar findings and trade-offs as we highlighted for agriculture.

In each context, a person's frame of reference, consisting of assumptions, values, norms, knowledge and interests, will be balanced differently, resulting in different prioritizations and selections of indicators and indicator sets (Te Velde et al. 2002). To build reliable sustainability assessments in the future, we have to recognize that definitions of sustainability, as well as the selection of indicators, vary with individual differences in context and perceptions (Bell and Morse 2008; Gasparatos 2010). Although each criterion could be considered valuable in its own sense, personal context may lead individuals to prioritize criteria differently. For example, one might value quantification but be willing to compromise that value for affordability purposes. This is not because affordability is considered more valuable than quantification, but results from a given context. The current ranking exercise deliberately asked respondents to order the criteria according to 'importance' for the sustainability assessments, knowing that 'importance' inevitably conflates practical constraints, systems understanding (i.e. what is most likely to affect sustainability outcomes) and, most of all, values of the assessors or subjects. This very conglomeration of multiple criteria is embedded in sustainability assessments and presents a challenge for achieving consensus, trust and collective action to transform agricultural systems to more sustainable practice.

Indicator selection is, however, not the only consideration in the design of sustainability assessment tools. Decisions on the purpose, assessment process (i.e. who measures and how), reporting and evaluation also influence the sustainability assessment (Binder et al. 2010). We urge sustainability researchers to more explicitly acknowledge that priorities and values play an important role in science, especially in sustainability assessments (Alrøe and Kristensen 2002; Alrøe et al. 2016). Describing and reflecting on the context of sustainability assessments, including prioritization of selection criteria, selection of indicators, methods and reference values, is crucial to improve the transparency and relevance of sustainable assessment tools.

4.3 Collaborative processes and participation as an answer to context specificity, plurality and flexibility

Dialogue is an important tool to improve transparency of prioritization and selection procedures to develop sustainability assessment tools. Stakeholder collaboration can support the development and dissemination of more robust conceptions of sustainability (De Mey et al. 2011; Bell and Morse 2008). Furthermore, selecting indicators can be seen

as a process of joint learning and knowledge development to help those involved to make decisions that enhance sustainable development of agriculture (Pretty 2008). The selection process stimulates stakeholders to recognize and accept their role and responsibility for achieving a more sustainable practice (Triste et al. 2014). Including all stakeholders may lead to wide-ranging goals, a broader focus and even inconclusiveness, which some professionals may find unsettling. It will also take much longer to establish monitoring and research because co-design and relationship building must occur first (Moller et al. 2009). Participatory co-development should, however, be seen as a crucial process for the interpretation and operationalization of sustainability (Owens 2003). Agrawal (2005) showed how involvement in monitoring itself has crucial roles in triggering individual transformation of the values and actions of environmental citizens. Mindful facilitation of such processes is considered necessary to create commitment, shared understanding and trust and to acknowledge and manage power asymmetries (Ansell and Gash 2008; Moller et al. 2009; Barnaud and Van Paassen 2013).

In summary, selection of indicators should be a process in which the stakeholders affected are involved, not just for the sake of participation (Bell and Morse 2008), but also to create relevant and context-specific assessments to improve sustainability performance (Binder et al. 2010; Gasso et al. 2015). Co-design and self-monitoring of indicators does much more than securing agreement and cooperation or reducing cost—it also requires a trigger for changing the orientations and actions of the participants, in this case mainly farmers, towards more environmentally caring outcomes. However, adoption of additional sustainability criteria from new participatory processes may be more difficult where market accreditation initiatives have already codified what must be included, especially by stipulating standards that must be met for gaining market access.

4.4 Future approaches and research for selecting sustainability indicators

As suggested by Bell and Morse (2008): 'Rapid and participatory tools for developing our thinking and modelling concerning measures of sustainability are of value to a wide range of stakeholders within development policy'. The ranking surveys provided such a tool and provoked discussions regarding the selection criteria for indicator selection within a broad group of sustainability experts. The results can be used to discuss ways to develop consensus with time through evolving views and concepts. The higher consensus score found in the NZSD ranking of selection criteria of individual indicators could be seen as an example of improved consensus resulting from close collaboration over years.

We hypothesized that experts with experience in researching agricultural sustainability and measuring the pressures, states and responses of agricultural systems would be more likely to find consensus on criteria for indicator selection than a group of more diverse stakeholders like growers, industry representatives, regulators and land use policy analysts, marketers and consumers. Our results need to be tested further to reveal if the apparent dilemma uncovered in our preliminary surveys can be generalized to other sustainability assessment frameworks. It would be particularly useful to deploy qualitative research methods to discover why experts prefer different types of indicators and indicator sets. Moreover, case studies of research and stakeholder collaborations within a defined context (e.g. region, assessment goal, end-users and priorities) could provide insight into how consensus develops.

To incorporate context specificity and flexibility, selection criteria as well as indicators and their reference values can be made context-specific. Furthermore, different sets of weights can be used to aggregate indicators in sustainability assessment tools. The same

indicators can then be measured by all decision-makers and subsequently (post hoc) aggregated and filtered to match different contexts and applications, i.e. each indicator is multiplied by a 'weight' ascribed by a given context or stakeholder (Cloquell-Ballester et al. 2006; Sadok et al. 2009; Elsaesser et al. 2015). There is a need to develop, test and cross-calibrate methods to measure weights ascribed by different participants to different indicators and sustainability dimensions.

Interactions between agricultural systems and their environment require thinking through different factors such as spatial and temporal scales, institutional behaviours, and knowledge types (Belt and Blake 2015). This requires the development of techniques for combining very different types of indicators (Alrøe et al. 2016). Different indicator types also relate to different functions of the sustainability assessment tools (Marchand et al. 2014). A possible solution to address context-specific needs is the development of modular tools through which end-users can select subsets of indicators within the sustainability assessment depending on the goal of the project and local conditions related, for example related to data availability (Marchand et al. 2014). Clearly, metrological research (the science of measurement) on what gets included and left out of multi-criteria sustainability assessments is an urgent priority. In this process, we can learn from other fields. For example, large-scale multi-criteria evaluation is used in biodiversity conservation to develop transparent traceability indicators (Ferraro and Pattanayak 2006).

Answers to several more overarching research questions could improve collaboration, trust and usability of future sustainability assessments and identify important features of the way they are designed and promulgated to accommodate widely different contexts, goals and values of the stakeholders. We here present useful research questions on the main themes of this paper:

Defining the sustainability construct:

- How important is it to define sustainability itself?
- Does a single definition lock in or exclude some participants, or does it provide clarity and unity of action?

Metrics of assessments:

- Do sound indicators recorded by farmers actually deliver the better farming outcomes they promise?
- How is a practice-based indicator system coherent with traditional farm management tools employed at the farm and is it possible to link these two 'entities' in an efficient way?
- What are robust scaling methods to reliably combine divergent indicators and measurement types at domain, outcome, objective and indicators levels?

Link between sustainability constructs and the metrics:

- Do we need to reach consensus on the importance of criteria for indicator selection?
- How much does our apparent divergence in what is important to measure in sustainability assessment reflect differences in values or in practical constraints?
- Might stakeholders find much stronger consensus and comparability around the higher-order goals, outcomes and objectives in the sustainability framework even if they differ sharply on how to measure performance at the base indicator levels?
- What would success look like (i.e. what difference does any sustainability assessment actually make to land, society, economy and governance and how would this be measured)?

Organizing the process in function of legitimacy, integrity, trust and outcomes:

- Does participation in designing the sustainability assessment tool, selecting indicators and design of their measures lead to substantive benefits for learning and transformation of values of farmers, or does it precipitate threats to the completeness, integrity and trust of the assessment by others?
- Does the very act of codifying sustainable practice and measuring it lead to participants focussing on the assessment itself rather than the ultimate goal of seeking more sustainable farming solutions?
- Do sustainability standards and market accreditation schemes provide adequate scope and rigour for whole systems assessment, or is there a need for complementary and supplementary assessments by independent civil agencies?
- Must such processes always be initiated from within a community of practice and, if so, should the sustainability assessment only aim for legitimacy within that same community?
- How can farmers' local knowledge be given legitimacy and voice alongside less situated knowledge of the type favoured by external experts and process professionals like policy makers, regulators and scientists?
- What are the most effective cross-scale bridging institutions and processes for linking distant stakeholders in food systems when opinions differ on the best way to measure it?

The TempAg research network seeks to compare sustainability performance of diverse OECD countries and agricultural sectors throughout temperate regions. The NZSD tools are largely designed to meet market verification needs, but also to become learning tools for farmers. Our initial survey warns that achieving consensus around sustainability assessment tools requires methods to account for participant variation. A single, universal sustainability assessment tool cannot be applied as a universal gold standard across communities and contexts. Harmonization of the overarching sustainability assessment framework, as done recently by FAO's SAFA programme (FAO 2013), is a huge step forward to drive consensus around criteria and procedures for sustainability assessments. However, what is left out of a sustainability assessment may be just as crucial as what is included or how it is measured. Our study suggests that considerable flexibility is needed in prioritizing indicators, that the stakeholders need to run their own process of selection and design of the indicators, and that this process be well documented so that robust analysis and comparisons can be made.

Finally, we wish to emphasize that although sustainability indicators are relevant to monitor sustainability performance, indicators reduce the complexity of a system into simplified measures. In-depth understanding of the sustainability of a system should therefore embrace a systems approach by addressing the context and interactions of systems (Schiere et al. 1999).

5 Conclusion

In this study, we found a lack of consensus amongst experts about what constitutes reliable knowledge and useable datasets for assessing sustainability. Although divergence of opinion in design criteria has been widely discussed, this first quantification of the degree of difference in opinion is startling and calls for a radical rethink about how

complementary approaches to sustainability assessments are used alongside each other to ensure a plurality of views and maximum collaboration and trust amongst stakeholders. We have to accept that people have different ways of assessing what is reliable knowledge and do so in a collaborative learning process. A useful start to a collaborative learning process is to recognize how sustainability is operationalized through scales and in different contexts. The process by which indicators and sustainability assessment tools are established may be a much more important determinant of their success than the final shape of the assessment tools. Such an emphasis on process would make assessments more transparent, transformative and enduring.

Acknowledgments We would like to thank all the participants from TempAg and NZSD for their participation. We were also grateful for guidance and permission from Paul Hansen and Franz Ombler for deploying the 1000Minds software. We would like to acknowledge Peter Groffman and the anonymous reviewers for their constructive suggestions on an earlier version of this paper. This is the first paper of the Resilient Agricultural Production Systems team of the Temperate Agriculture Research Network and international collaboration initiated by OECD's Global Science Forum. NZSD's participation is funded by New Zealand's Ministry for Business, Innovation and Employment (contract AGRB1201).

References

Agrawal, A. (2005). *Environmentality. Technologies of government and the making of subjects*. Durham and London: Duke University Press.

Alrøe, H. F., & Kristensen, E. S. (2002). Towards a systemic research methodology in agriculture: Rethinking the role of values in science. *Agriculture and Human Values, 19*(1), 3–23.

Alrøe, H. F., Moller, H., Læssøe, J., & Noe, E. (2016). Opportunities and challenges for multicriteria assessment of food system sustainability. *Ecology and Society*. doi:10.5751/ES-08394-210138.

Ansell, C., & Gash, A. (2008). Collaborative governance in theory and practice. *Journal of Public Administration Research and Theory, 18*(4), 543–571. doi:10.1093/jopart/mum032.

Bäckstrand, K. (2004). Scientisation vs. civic expertise in environmental governance: Eco-feminist, eco-modern and post-modern responses. *Environmental Politics, 13*(4), 695–714. doi:10.1080/0964401042000274322.

Barnaud, C., & Van Paassen, A. (2013). Equity, power games, and legitimacy: Dilemmas of participatory natural resource management. *Ecology and Society*. doi:10.5751/ES-05459-180221.

Bell, S., & Morse, S. (2008). *Sustainability indicators: Measuring the immeasurable?*. London: Earthscan.

Belt, M., & Blake, D. (2015). Mediated modeling in water resource dialogues connecting multiple scales. *JAWRA Journal of the American Water Resources Association*, *51*(6), 1581–1599.

Binder, C. R., Feola, G., & Steinberger, J. K. (2010). Considering the normative, systemic and procedural dimensions in indicator-based sustainability assessments in agriculture. *Environmental Impact Assessment Review, 30*(2), 71–81.

Bockstaller, C., Feschet, P., & Angevin, F. (2015). Issues in evaluating sustainability of farming systems with indicators. *OCL Oilseeds and Fats, Crops and Lipids*. doi:10.1051/ocl/2014052.

Bockstaller, C., Guichard, L., Keichinger, O., Girardin, P., Galan, M. B., & Gaillard, G. (2009). Comparison of methods to assess the sustainability of agricultural systems. A review. *Agronomy for Sustainable Development, 29*(1), 223–235.

Cloquell-Ballester, V.-A., Cloquell-Ballester, V.-A., Monterde-Díaz, R., & Santamarina-Siurana, M.-C. (2006). Indicators validation for the improvement of environmental and social impact quantitative assessment. *Environmental Impact Assessment Review, 26*(1), 79–105. doi:10.1016/j.eiar.2005.06.002.

Dale, V. H., & Beyeler, S. C. (2001). Challenges in the development and use of ecological indicators. *Ecological Indicators, 1*(1), 3–10.

De Mey, K., D'Haene, K., Marchand, F., Meul, M., & Lauwers, L. (2011). Learning through stakeholder involvement in the implementation of MOTIFS: An integrated assessment model for sustainable farming in Flanders. *International Journal of Agricultural Sustainability, 9*(2), 350–363.

de Olde, E. M., Oudshoorn, F. W., Sørensen, C. A. G., Bokkers, E. A. M., & de Boer, I. J. M. (2016). Assessing sustainability at farm-level: Lessons learned from a comparison of tools in practice. *Ecological Indicators, 66*, 391–404. doi:10.1016/j.ecolind.2016.01.047.

Dovers, S. (2005). Clarifying the imperative of integration research for sustainable environmental management. *Journal of Research Practice, 1*(2), 1–19.

Elsaesser, M., Jilg, T., Herrmann, K., Boonen, J., Debruyne, L., Laidlaw, A. S., et al. (2015). Quantifying sustainability of dairy farms with the DAIRYMAN sustainability-index. In: *Paper presented at the European Grassland Federation*, Wageningen, The Netherlands.

FAO. (2013). *Sustainability assessment of food and agriculture systems (SAFA): Guidelines, version 3.0*. Rome: Food and Agricultural Organization of the United Nations.

Ferraro, P. J., & Pattanayak, S. K. (2006). Money for nothing? A call for empirical evaluation of biodiversity conservation investments. *PLoS Biology, 4*(4), e105. doi:10.1371/journal.pbio.0040105.

Gasparatos, A. (2010). Embedded value systems in sustainability assessment tools and their implications. *Journal of Environmental Management, 91*(8), 1613–1622. doi:10.1016/j.jenvman.2010.03.014.

Gasso, V., Oudshoorn, F. W., de Olde, E., & Sørensen, C. A. G. (2015). Generic sustainability assessment themes and the role of context: The case of Danish maize for German biogas. *Ecological Indicators, 49*, 143–153. doi:10.1016/j.ecolind.2014.10.008.

Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric statistical inference*. Berlin: Springer.

Hansen, J. (1996). Is agricultural sustainability a useful concept? *Agricultural Systems, 50*(2), 117–143.

Herzog, F., Balázs, K., Dennis, P., Friedel, J., Geijzendorffer, I., Jeanneret, P., et al. (2012). *Biodiversity indicators for European farming systems: A guidebook*: Forschungsanstalt Agroscope Reckenholz-Tänikon ART.

IAASTD. (2009). Agriculture at a crossroads: Synthesis report. International Assessment of Agricultural Knowledge, Science and Technology for Development (IAASTD) *Science and Technology for Development*. Island Press.

Jones, C., Cowan, P., & Allen, W. (2012). *Setting outcomes, and measuring and reporting performance of regional council pest and weed management programmes. Guidelines and resource materials*. Landcare Research Contract Report LC144: Landcare Research New Zealand Ltd.

Kates, R. W., Clark, W. C., Corell, R., Hall, J. M., Jaeger, C. C., Lowe, I., et al. (2001). Environment and development: Sustainability science. *Science, 292*(5517), 641–642. doi:10.1126/science.1059386.

Kendall, M. G., & Smith, B. B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics, 10*(3), 275–287.

Keulen, H. V., van Ittersum, M., & Leffelaar, P. (2005). Multiscale methodological framework to derive criteria and indicators for sustainability evaluation of peasant natural resource management systems. *Environment, Development and Sustainability, 7*(1), 51–69.

Komiyama, H., & Takeuchi, K. (2006). Sustainability science: Building a new discipline. *Sustainability Science, 1*(1), 1–6. doi:10.1007/s11625-006-0007-4.

Lebacq, T., Baret, P. V., & Stilmant, D. (2013). Sustainability indicators for livestock farming. A review. *Agronomy for Sustainable Development, 33*(2), 311–327.

Lee, W., McGlone, M., & Wright, E. (2005). *Biodiversity inventory and monitoring: A review of national and international systems and a proposed framework for future biodiversity monitoring by the Department of Conservation*. Landcare Research Contract Report LC0405/122.

Lupia, A. (2013). Communicating science in politicized environments. *Proceedings of the National Academy of Sciences, 110*(Supplement 3), 14048–14054. doi:10.1073/pnas.1212726110.

Marchand, F., Debruyne, L., Triste, L., Gerrard, C., Padel, S., & Lauwers, L. (2014). Key characteristics for tool choice in indicator-based sustainability assessment at farm level. *Ecology and Society*. doi:10.5751/ES-06876-190346.

Merfield, C., Moller, H., Manhire, J., Rosin, C., Norton, S., Carey, P., et al. (2015). Are organic standards sufficient to ensure sustainable agriculture? Lessons from New Zealand's ARGOS and Sustainability Dashboard projects. *Sustainable Agriculture Research, 4*(3), p158.

Moller, H., & MacLeod, C. J. (2013). *Design criteria for effective assessment of sustainability in New Zealand's production landscapes*. (Vol. 13/07, pp. 73): NZ Sustainability Dashboard Research Report.

Moller, H., O'Blyver, P., Bragg, C., Newman, J., Clucas, R., Fletcher, D., et al. (2009). Guidelines for cross-cultural participatory action research partnerships: A case study of a customary seabird harvest in New Zealand. *New Zealand Journal of Zoology, 36*(3), 211–241. doi:10.1080/03014220909510152.

Niemeijer, D., & de Groot, R. S. (2008). A conceptual framework for selecting environmental indicator sets. *Ecological Indicators, 8*(1), 14–25. doi:10.1016/j.ecolind.2006.11.012.

OECD. (2001). *Environmental indicators for agriculture. Methods and results* (Vol. 3). Paris: Organisation for Economic Co-operation and Development.

Ostrom, E. (2009). A general framework for analyzing sustainability of social-ecological systems. *Science, 325*(5939), 419–422. doi:10.1126/science.1172133.

Owens, S. (2003). Is there a meaningful definition of sustainability? *Plant Genetic Resources: Characterization and Utilization, 1*(01), 5–9.

Parris, T. M., & Kates, R. W. (2003). Characterizing and measuring sustainable development. *Annual Review of Environment and Resources, 28*, 559–586.

Popa, F., Guillermin, M., & Dedeurwaerdere, T. (2015). A pragmatist approach to transdisciplinarity in sustainability research: From complex systems theory to reflexive science. *Futures*. doi:10.1016/j.futures.2014.02.002.

Pretty, J. (2008). Agricultural sustainability: Concepts, principles and evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1491), 447–465.

Pretty, J., Sutherland, W. J., Ashby, J., Auburn, J., Baulcombe, D., Bell, M., et al. (2010). The top 100 questions of importance to the future of global agriculture. *International Journal of Agricultural Sustainability, 8*(4), 219–236.

Reed, M. S., Fraser, E. D. G., & Dougill, A. J. (2006). An adaptive learning process for developing and applying sustainability indicators with local communities. *Ecological Economics, 59*(4), 406–418.

Sadok, W., Angevin, F., Bergez, J.-E., Bockstaller, C., Colomb, B., Guichard, L., et al. (2009). MASC, a qualitative multi-attribute decision model for ex ante assessment of the sustainability of cropping systems. *Agronomy for Sustainable Development, 29*(3), 447–461. doi:10.1051/agro/2009006.

Schader, C., Grenz, J., Meier, M. S., & Stolze, M. (2014). Scope and precision of sustainability assessment approaches to food systems. *Ecology and Society*. doi:10.5751/ES-06866-190342.

Schiere, J. B., Lyklema, J., Schakel, J., & Rickert, K. G. (1999). Evolution of farming systems and system philosophy. *Systems Research and Behavioral Science, 16*(4), 375–390.

Seimon, A., Plumptre, A. J., & Watson, J. E. M. (2012). *Building consensus on Albertine Rift climate change adaptation for conservation: A report on 2011–2012 workshops in Uganda and Rwanda*. WCS Workshop Report. New York, USA: Wildlife Conservation Society (WCS).

Seimon, A., Yager, K., Seimon, T., Schmidt, S., Grau, A., Beck, S., et al. (2009). Changes in biodiversity patterns in the high andes—Understanding the consequences and seeking adaptation to global change. *Mountain Forum Bulletin, 9*, 25–27.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York, NY: McGraw Hill.

Sommerville, M. M., Milner-Gulland, E., & Jones, J. P. (2011). The challenge of monitoring biodiversity in payment for environmental service interventions. *Biological Conservation, 144*(12), 2832–2841.

Steinfeld, H., Gerber, P., Wassenaar, T., Castel, V., Rosales, M., & Haan, C. D. (2006). *Livestock's long shadow: Environmental issues and options*. Rome: Food and Agriculture Organization of the United Nations (FAO).

Te Velde, H., Aarts, N., & Van Woerkum, C. (2002). Dealing with ambivalence: Farmers' and consumers' perceptions of animal welfare in livestock breeding. *Journal of Agricultural and Environmental Ethics, 15*(2), 203–219. doi:10.1023/A:1015012403331.

Triste, L., Marchand, F., Debruyne, L., Meul, M., & Lauwers, L. (2014). Reflection on the development process of a sustainability assessment tool: Learning from a Flemish case. *Ecology and Society*. doi:10.5751/ES-06789-190347.

Yager, K., Ulloa, D., & Halloy, S. (2009). *Chapter 16. Conducting an interdisciplinary workshop on climate change: Facilitating awareness and adaptation in Sajama National Park, Bolivia.* (Interdisciplinary Aspects of Climate Change). Hamburg: Hamburg University of Applied Sciences.