Australasian Marketing Journal ■■ (2017) ■■–■■

Contents lists available at ScienceDirect

Australasian Marketing Journal

journal homepage: www.elsevier.com/locate/amj

How small sample size and replication can increase accuracy in experiments: Lessons that marketing may learn from agricultural scientific method

Robert Hamlin *

Department of Marketing, University of Otago, PO Box 56, Dunedin 9016, New Zealand

ARTICLE INFO

Article history:
Available online

Keywords:
Marketing
Science
Experiment
Sample size
Replication

ABSTRACT

This paper examines the use of small sample sizes and replication in marketing experimentation, including full factorials, fractional factorials, Latin squares and their derivatives such as conjoint analysis. It is well understood within agricultural research that the sample size used within these experiments should be kept to a minimum if maximum reliability is to be achieved. This understanding, which underlies the massive success of agricultural research in the last century, does not appear to have been transferred to marketing. This article explains the logic behind this counterintuitive claim. It then discusses the links between the use of small sample size and replication in experimental research. It concludes that the current very low level of replication in marketing can be related to a very basic mismatch between academic marketing's theoretical expectations of replication outcomes and the degree to which these expectations can be meaningfully achieved by replication within any living environment.

1. Introduction

This article examines the apparently counter-intuitive claim that smaller sample sizes give more accurate results when combined with certain experimental techniques that are very popular in academic marketing.

The advantages of a small sample size are well understood in agricultural science, where these techniques originated, and this realisation forms a basic tenet of agricultural science research method. The article uses agricultural research situations to demonstrate why this tenet is both valid and important. It also demonstrates how the philosophy, assumptions and methodologies that underlie this tenet also lead to a greater requirement for replication as a routine part of hypothesis testing, both within individual research exercises (intra-research replication) and between them as part of subsequent published discourse (inter research replication).

There appears to be no significant understanding of this crucial methodological tenet within the academic marketing literature, which displays no recognition of the advantages of small sample sizes, but does display a recognised deficiency in its rate of replication (Easley and Madden, 2013; Easley et al., 2000). The means by which small sample size tenet can then be effectively applied to marketing research are therefore demonstrated with a short theoretical discussion and a single worked example.

The article concludes with a discussion on the relationship between the use of small sample sizes and the use of replication in research. It is proposed that the widely reported lack of replication in research is significantly associated with a single underlying epistemological cause: namely academic marketing's general aversion to the use of small samples, and to the related lack of a widespread realisation as to why the research environment within which the discipline operates makes it advantageous to use large numbers of small sample intra and inter research replication exercises when using some of its most popular experimental techniques.

Previous commentaries on replication in marketing have focussed upon the distinction between replication practice within marketing as a 'social' science and replication within the 'natural' or 'life' sciences (Easley et al., 2013). Agricultural science and the approaches that have led to its remarkable success over the last eighty years have received little attention. This is unfortunate because agricultural research is a 'hybrid' discipline with a strong commercial and applied emphasis, which shares many similarities with both the natural and marketing sciences. It has also historically been a major technical donor to the social science research toolbox. The pragmatic solutions that it has discovered and successfully applied to its research processes as it continues to feed the World are highly applicable to the research issues that marketing faces.

* Fax: +64 3479 8172.
E-mail address: rob.hamlin@otago.ac.nz.

http://dx.doi.org/10.1016/j.ausmj.2017.05.002
1441-3582/

2. Agricultural science's contribution to marketing experimental technique and method

Marketing utilises a variety of experimental designs that were initially developed for use in agricultural research (Banks, 1965; Brunk and Federer, 1953a, 1953b; Cox, 1964). These designs have come to Marketing either directly by transfer (Latin square, fractional factorial, full factorial) or by subsequent development of these initial designs (conjoint analysis) (Hamlin, 1997, 2005). With one or two caveats, the transfer has been a happy one (Hamlin, 2005). However there are several important insights derived from their development and widespread successful application within Agriculture that do not seem to have been sufficiently widely disseminated within the marketing research literature.

Exactly why these insights have failed to transfer between the disciplines can only be a matter of historical conjecture. However, the author's own extensive experience as both a commercial/academic agricultural researcher and an academic marketing researcher may provide a possible cause. Agricultural researchers tend to apply only a very small and stable suite of closely related experimental techniques that have been formalised for a very long time. The nature of the research environment means that they also execute them at a very high volume and either apply or publish the results very quickly (Gomez and Gomez, 1984; Sunding and Zilberman, 2001). This creates a 'density' of experience within the community that academic marketing, with its larger toolbox, greater diversity and slower research cycle rate, simply cannot match.

As a consequence published and thus visible experimental 'techniques' are supported by a much larger, unpublished, orally transmitted, invisible, but nonetheless highly critical body of experimental 'method' within agricultural research. This detailed experimental method is rarely written up for publication as it is bulky and common inter-researcher expertise is assumed. It thus cannot be easily referenced by an 'outsider', and therefore can only be transmitted to another discipline by a significant transfer of personnel. Thus its absence within academic marketing is understandable.

This article examines the nature and implications of just one of these agricultural research method-derived insights: That it is a key requirement of experimental reliability that results are derived from individual samples that are as *small* as possible. Small size is achieved by using an experimental pattern that has the highest efficiency with regard to the individual treatment conditions that are required, and by using the smallest possible sample size within each treatment condition. The stability and generalisability of any conclusion is tested by systematic replication of these small exercises. These replications are commonly both intra-study (to test stability) and inter-study (to test generalisability) to the method and conclusions of any one published report.

3. The economic and practical advantages of small sample size

High efficiency in an experimental design has the obvious attraction that a result can be obtained after a much lower expenditure of time, money and other research resources. The same comments can be made with regard to a small individual sample for each treatment condition within any such design. A further benefit of both of these features is that any experiment that possesses them may be administered with a very much lower degree of disruption of the environment in which it is undertaken.

This is important as much of the research work using agricultural designs since their introduction to marketing in 1953 has been administered in difficult to access field environments, such as retail stores or supermarkets (Brunk and Federer, 1953a, 1953b; Cox, 1964; Dodds et al., 1991; Kennedy, 1970; Montaguti et al., 2015; Orth and Malkewitz, 2012; Rui and Meyers-Levy, 2009). Under such circumstances, where the co-operation of a commercial partner is required, the efficiency of the experimental design may determine if consent to conduct field research is granted at all.

4. The technical advantages of small sample size – Fisher's two principles

Beyond these advantages there is a much more subtle, yet highly important benefit endowed by high efficiency. Nearly all the experimental designs sourced from agriculture are instruments of *parallel* comparison, which rely on the controlled application combinations of the independent variables to *equivalent* experimental units. The mean responses of these individual units are then compared to the mean response of the entire experimental population, or to a single nominated 'control' condition if a partially confounded (fractional factorial) design is being used. The main and non-additive effects of the controlled independent variables are then deduced algebraically from the deviations of individual conditions from the population mean or a nominated control condition. Simple statistical tests such as ANOVA are then used to test the stability of these algebraic manipulations.

The larger the experiment becomes in terms of the number and/or size of the individual experimental units deployed, the harder it becomes to either ensure or reasonably assume that these units are all either internally homogeneous or equivalent to each other for the purposes of these comparisons. The effects of the controlled independent variables will be increasingly moderated by other non-controlled variables that are unavoidably present within the sample environment. As the sample environment increases in size, the greater the chance becomes that effects of these non-controlled external variables will not be uniform, either within or between individual samples.

The less certain the equivalence of the treatments is, then the less reliable the results of the overall experiment will be. It is for this reason that Sir Ronald A. Fisher, the initial developer of nearly all these experimental designs and related statistical tests for agricultural purposes (Fisher, 1925, 1935), made the following direct comment on experimental method in the form of two principles:

> "… the problem of designing economical and effective field experiments is reduced to two main principles (i) the division of the experimental area into the plots as small as possible …; (ii) the use of [experimental] arrangements which eliminate a maximum fraction of soil heterogeneity, and yet provide a valid estimate of residual errors."
>
> (Fisher, 1950, p. 510)

As Fisher does not do so, and because the issues relating to them are easier to demonstrate in this way, it is necessary to elaborate upon and demonstrate these two principles in their original agricultural experimental method context before their significance for marketing situations can be discussed. It is acknowledged within agriculture that there is no such thing as an entirely homogeneous environment, outside of a hydroponic cell. While experiments can be conducted within hydroponic cells, the conditions within them are so far removed from the reality 'in the ground' that attempts to extrapolate results from them into the more general environment have to be treated with considerable caution. The situation is a close analogue to the lab experiments that are frequently published in the marketing literature (Calder et al., 1981; Koschate-Fischer and Schandelmeier, 2014).

As a consequence most agricultural field experiments are exactly that; they occur in a field, and that field, even if it is an open flat block in Kansas, will have a range of uncontrolled environmental conditions existing within it (Fig. 1). The precise nature of these

Please cite this article in press as: Robert Hamlin, How small sample size and replication can increase accuracy in experiments: Lessons that marketing may learn from agricultural scientific method, Australasian Marketing Journal (2017), doi: 10.1016/j.ausmj.2017.05.002

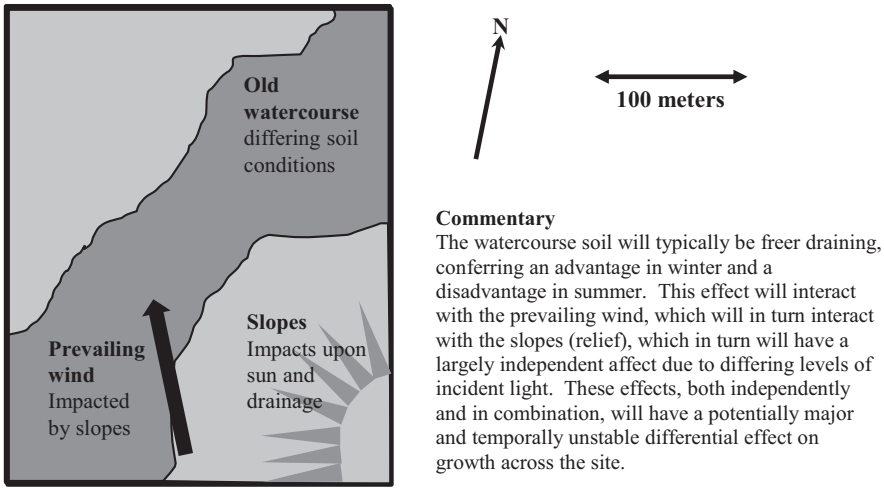

Fig. 1. A typical agricultural field research environment.

variations and their impact upon the experiment, individually and in combination with one another, will not be fully known to the researcher. Using the field environment shown in Fig. 1, Fig. 2 shows why Fisher identified the principle (*i*) that individual plot sizes have to be as small as possible. In both Fig. 2a and b we have a 3×3 Latin square design with nine plots. The design is capable of handling three independent variables. However, in Fig. 2a those plots are approximately 10×10 meters, while in Fig. 2b they are approaching 80×80 meters.

The larger experiment confers some advantage in masking very local variations in growth conditions, such as local parasites and 'edge effects' (Langton, 1990), but the enlarged design incorporates both the watercourse and the slope. In the smaller design, growing conditions for all plots are broadly comparable, but the larger design has plots in both extremely favourable and unfavourable locations. As the analysis of variance of the experiment is blind to these uncontrolled variations, they will appear in the analysis of variance table for the larger design in the form of an increased error estimate, an erroneous assignation of main effects for the three controlled variables, or both. A 'Type II' error is a more likely outcome of the apparently more imposing exercise shown in Fig. 2b than it is for Fig. 2a.

Fisher's second principle (*ii*) that the use of [experimental] arrangements which eliminate a maximum fraction of soil heterogeneity, and yet provide a valid estimate of residual errors, can be explained using the same example environment. The Latin square design in Fig. 2a is capable of identifying the nature and scale of the main effects of three independent variables, but only in the absence of non-additivity. If the researcher chooses to set up their independent variables in a manner that makes non-additivity between them a potential issue, then they may decide to use a larger and more 'powerful' experimental design, such as a full factorial, that can establish whether non-additivity is present or not.

However, this increase in size comes with a cost, as Fig. 2c shows. Here we have the 3×3 Latin square expanded out to a $3 \times 3 \times 3$ full factorial with the same smaller plot size. The larger design, anchored on the same North West point, now extends into the dry watercourse with 13 of its 27 plots subjected to its very different growing conditions. Once again the design and the analysis of it are blind to this new and potentially substantial uncontrolled input. The more complex design and the randomised plot location within it makes a systematic bias from this source less likely, but there is a high probability of a Type II error with regard to either the main effects or the interactions between them.

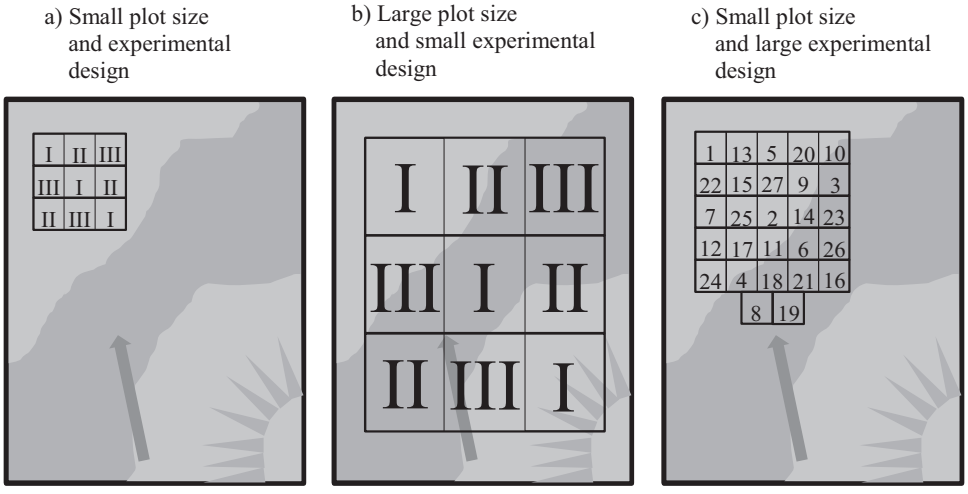

Fig. 2. The impact of larger design on sample homogeneity in a typical agricultural research experimental environment.

Please cite this article in press as: Robert Hamlin, How small sample size and replication can increase accuracy in experiments: Lessons that marketing may learn from agricultural scientific method, Australasian Marketing Journal (2017), doi: 10.1016/j.ausmj.2017.05.002

5. Small sample sizes and replication

An obvious response to the argument for smaller experimental sample populations is that it is as possible for a small experimental population to lie across a change in environment as a large one, and this is certainly true (Fig. 3a). However, the chances of this happening are reduced if a smaller experimental population is used, even if the location is random. If the locations are deliberately selected by the researcher to minimise environmental variations within the plots, then a smaller sample will increase the opportunities that are available to achieve this within any particular environment.

In addition, the environmental map in Fig. 3a is a simplification of reality which shows only gross discontinuities in the environment. Between these specific discontinuities progressive changes will still occur within each region (Fig. 3b), the impacts of which will progressively impact upon the error within each experiment as its size increases, leading to a progressive and unavoidable chance of a type II error as the plot size increases.

This leads on to a second issue, that of external validity. There are very obvious limitations to the external validity of any research that draws its experimental population from a single environmental condition when other environments to which any conclusions might be applied have the potential to vary widely beyond this state. As there is no trade-off between internal and external validity in any experimental situation (any reduction in internal validity leads to a concomitant reduction in any external validity), the answer to this is to aim for maximum internal validity for any individual experimental exercise by the use of small, homogeneous samples, and to replicate the exercise in other environments as many times as is required to establish the extent to which any generalisations may hold good, and to which the findings may therefore logically be applied. Such an arrangement is shown in Fig. 3c. The arrangement minimises within experimental variation, and would be achieved by careful survey prior to laying out the sites.

The analysis of variance can be set up using the same data to treat these three exercises as independent replications on three tables, or as one table with either the replication as an independent variable or with the replications as an integral replication factor. However, many agricultural researchers, in commercial research situations at least, would treat the individual experiments as individual replications with separate tables, and would analyse and examine the individual outcomes 'cognitively' before applying any other overarching statistical tests to them.

The outcome of the exercise in Fig. 3c may be that, within this environment at least, the result is stable across all three exercises. It may also be that each exercise returns a radically different result in each location, in which case the conclusion may be that there are main effects present, but that they are too environmentally labile to be considered to represent a generalisable conclusion, and that in future local experimentation would have to be conducted in order to establish a local outcome, both in this environment and anywhere else.

That an outcome was observed in each exercise, but that it is not generalisable across replications, is a valuable finding in itself, and would be more useful than the 'no observable main effects' type II error that would be the likely outcome of either of the two larger experimental designs shown in Fig. 2b and c if they had been undertaken in a similar location-labile research situation. For this reason, full replication is routinely undertaken as an integral part of any individually reported research exercise in agricultural science (e.g. Jones et al., 2012), a situation that is rare in equivalent marketing exercises, which have a tendency to report on a series of different research exercises (e.g. Olson et al., 2016).

Clearly the array of environments and objectives that may be encountered by a scientist pursuing this type of research is virtually infinite, and the laying out of sites to achieve the minimum within experimental sample variation involves the consideration of such a large number of potentially interacting environmental variables that it becomes as much an art form as a procedure. It is an art form that is transmitted within the discipline, and to this researcher, by interpersonal verbal means, rather than by the written word.

This issue is illustrated in Fig. 4. In Fig. 4a we have the layout in Fig. 3c reproduced. The researcher has used their (informed!) judgement to identify the three sites shown, and has set up an integrated replication on each of them. It might seem better practice to randomise these replications within a single design in order to avoid any systematic bias in the replications, producing the designs shown in Fig. 4b and c. If this approach is pursued, then the uncontrolled environment related error produced by the additional replications will be randomised, but not eliminated, and will thus end as a general error term. Given this outcome, the designs in Fig. 4b and c

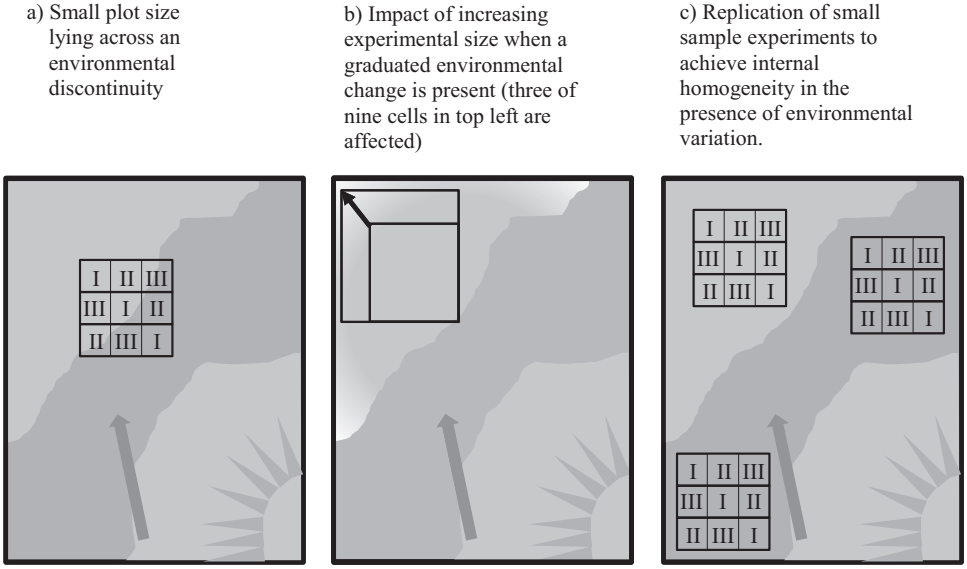

Fig. 3. Internal validity, external validity and replication.

Please cite this article in press as: Robert Hamlin, How small sample size and replication can increase accuracy in experiments: Lessons that marketing may learn from agricultural scientific method, Australasian Marketing Journal (2017), doi: 10.1016/j.ausmj.2017.05.002

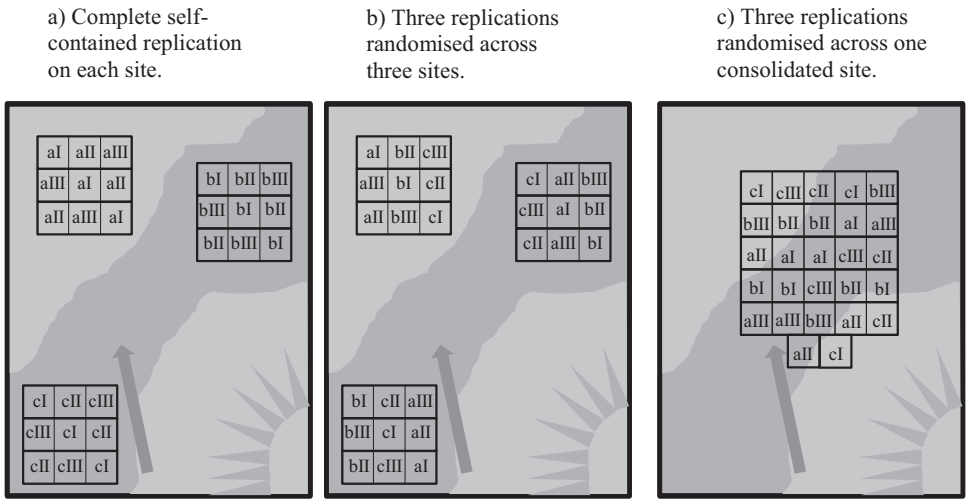

Fig. 4. Different approaches to replication.

are equally uninformative. By contrast the systematic portion of the variation generated by the replications in design in Fig. 4a will be partitioned as the sum of squares for replications, thus keeping it out of the error term and reducing the odds of a type II error.

If design in Fig. 4a has been used, a significant result for replications can also be studied by the researcher in order to come up with some explanation for it in terms of the known or discovered environmental variations between each self-contained replication. The cost of this is that the non-random nature of the design in Fig. 4a has to be described and justified, as does any non-statistical theoretical explanation of the outcomes of it. This might make a researcher aiming for an academic publication think twice about doing it. A commercial researcher, who has no such worries, may well consider it a price worth paying for the additional information and opportunity for discovery.

Perhaps in order to avoid the admission of the role of subjective judgement in experimental design and analysis, and in order to make his recommendations as usefully applicable as possible, Fisher restricted his comments on experimental method to the two principles noted previously, which in turn may be reduced to a single dictum: *'Within the requirements of internal validity for any individual research situation, then the smaller the sample can be made, the better'*.

6. Technical implications

The issue of selecting an optimal sample size can also be expressed as a trade-off within the analysis of variance table (Fig. 5). Assuming all main effects are stable, then increasing the sample size of an experiment can lead to an increase in its accuracy/power that is statistically predictable via decreases in the mean squared error term (MSE) that acts as the denominator for the 'F' ratio. This reduction in MSE is achieved either by decreasing the numerator, the total sum squares for error (SSE), by increasing the stability of individual 'cell' observations by increasing the number of data points per cell, or by increasing the denominator term, the numbers of degrees of freedom (df) associated with the sum of squares for error, by increasing the number of 'cells' in the experiment. This is a predictable relationship that is expressed by the dashed line in Fig. 5. This relationship takes no account of extraneous environmental variations within the sample.

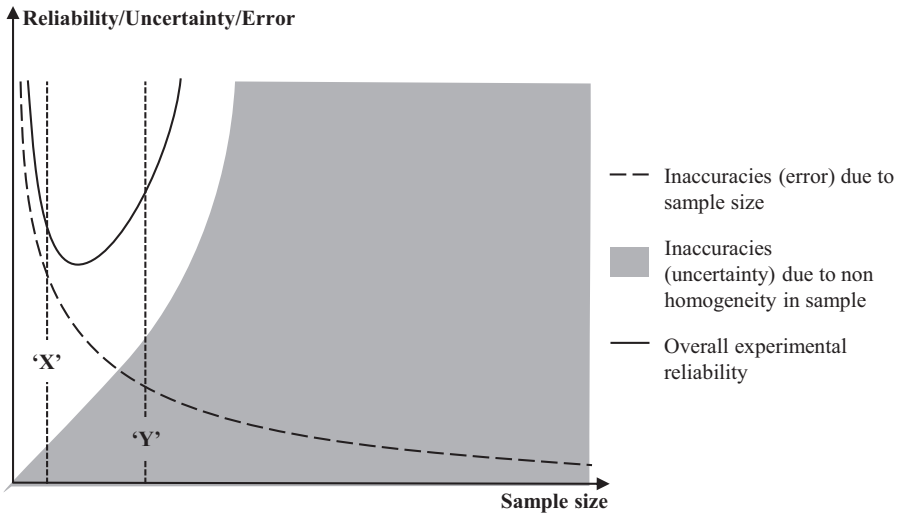

Fig. 5. The 'optimal' sample size for an instrument of parallel comparison.

Please cite this article in press as: Robert Hamlin, How small sample size and replication can increase accuracy in experiments: Lessons that marketing may learn from agricultural scientific method, Australasian Marketing Journal (2017), doi: 10.1016/j.ausmj.2017.05.002

However as the size of an experiment relying on parallel comparison increases in size it becomes increasingly subject to external sources of uncertainty due to the potential presence of unknown/uncontrolled non-homogeneity in the experimental environment. Unlike the potential sources of error associated with a smaller sample size, the extent and nature of the distortions introduced by these external factors of uncertainty cannot be statistically estimated or related to any specific sample size. It is therefore a true uncertainty rather than a statistically quantifiable risk and is thus (imperfectly) presented as a shaded area in the graph in Fig. 5.

If the assumption that main effects are stable within the sample is violated, and the experiment has been properly randomised, then this variation also will express itself as an increase in the error term as the sample size increases as a part of the same process, thus shifting the shaded area to the left and an increase in the likelihood of a Type II error. If the experiment has not been properly randomised with its environment within each experimental exercise, then there is the rather more unfortunate possibility that the same process will generate a Type I error.

The fact that this second source of error is an unquantifiable uncertainty, rather than a chance of error that may be estimated statistically, means that researchers should 'hug' the right hand side of the dashed error line in Fig. 5, and should make their experiments as small as possible within the constraints of the relationships that they are investigating. This is shown by the vertical line 'X' in Fig. 5. In Fisher's terms 'X' is the minimum point at which the researcher can '*provide a valid estimate of residual errors*'. The sample required to achieve 'X' may be minimised by either decreasing the number of observations per cell to a minimum or by reducing the number of cells required. This second requirement often provides greater capacity for economy. It involves a ruthless examination of the planned outcomes of an experiment, and the elimination of any hypothesis test that increases the required sample size, but that is not absolutely essential.

Adopting the most efficient experimental design that satisfies the minimum research requirements allows the researcher to reduce the required overall sample size, and thereby reduce the chance of an uncertain distortion occurring due to a non-homogeneity within the experimental environment. Such a smaller experiment may be narrower and statistically less powerful, but at least it is predictably so. Yielding to the temptation to use a larger sample size to increase the breadth, power and perceived reliability of an experiment risks the outcome illustrated by the line 'Y' in Fig. 5, where an increased sample size, with its attendant costs, appears to reduce the risk, but actually reduces the overall reliability of the outcome relative to the smaller sample size 'X'.

Agricultural researchers thus use small individual samples, not to save money, but to increase reliability for the reasons stated above. An approach demonstrated by Berzsenyi et al. (2000) (Fig. 6). As in the last seventy years commercial agriculture has achieved the unlikely feat of feeding the World in the face of a nearly five-fold increase in its population, largely via a 'green revolution' that has been informed by routine generation of scientific research results based upon the small sample size and routine internal and external replication tenets, this particular aspect of their approach to experimental design and execution are to be respected by those who seek to use similar designs within their own discipline.

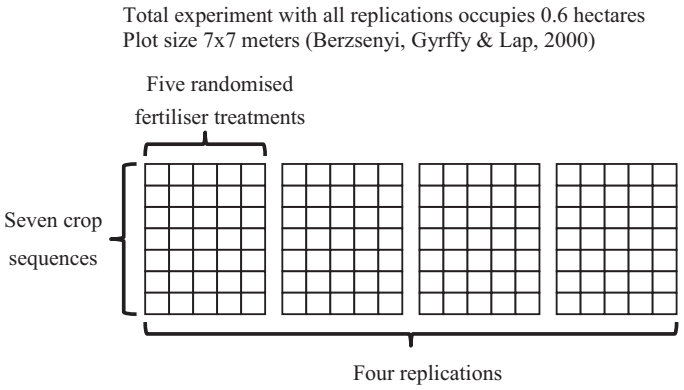

Fig. 6. Small sample size as applied in agriculture.

7. Applying Fisher's principles to marketing: the process

Two of the advantages of agricultural research situations are that they are both simpler and extremely 'visual'. Environmental 'truths' that are immediately self-evident when examining a field of corn are not so when examining community of human beings. Samples are also fully under the control of the researcher and thus are available for a prolonged period. As a consequence the agricultural researcher can study an environment at their leisure before they commit to sampling, and at least some of its more significant variations will become manifest as a result. Social scientists such as marketers are not nearly so fortunate. The unit that they have to contend with is not the field plot or a conveniently captive and controlled group of animals, but a transient sample of humanity who must be persuaded to participate in the research as a minor part of their everyday lives that invariably lie completely beyond the control of the researcher.

While this environment may make it harder for researchers to immediately perceive the logic of Fisher's principles, they still apply to the social science situation to at least the same degree as they do to agriculture. They can thus be used to improve the reliability of marketing research if they are applied appropriately.

Fig. 7 takes the environment shown in Figs. 1 and 3b and directly reascribes social rather than agricultural variables to it. The 'space' in this example is not a specific geographic location, but an equivalent human consumer recruitment situation. Some would describe this graphical portrayal of a group of humans as a gross oversimplification, and they would be right to do so. There are many more variables within this population, other than those shown here, that might significantly influence both each other and any independent variables that the researcher might choose to apply to the sample represented by the square in Fig. 7.

It may well be impossible for a researcher to identify all of these variables, when trying to create a homogeneous sample for an experiment, and thus they face an issue of uncertainty rather than risk. Such invisible uncertainty lies behind all the seemingly 'concrete' estimates of main effects and their statistical significance that are the published output of research of this type in the academic marketing literature. The degree to which it does so depends upon the approach taken by the researcher.

Given this, the logical response is to conform to Fisher's principles and seek to recruit a sample that is as homogeneous as possible with regard to as many variables as they can reasonably achieve, and then to refine the experimental design with a view to reducing the number of cells required for the experiment, and the number of individuals contributing to each cell. The use of the smallest sample possible reduces the probability that an uncontrolled environmental variable within the smaller sample will impact significantly upon the effects and interactions of the chosen independent variables to its minimum value.

It does not, however, eliminate it. As a consequence this small design is then replicated as many times as necessary to establish that stability has been achieved, and that an adequately reliable result for that particular environment can therefore be reported – or not, as the case may be. Such a combined result should have a very high degree of internal validity. If extension of the specific result

Please cite this article in press as: Robert Hamlin, How small sample size and replication can increase accuracy in experiments: Lessons that marketing may learn from agricultural scientific method, Australasian Marketing Journal (2017), doi: 10.1016/j.ausmj.2017.05.002

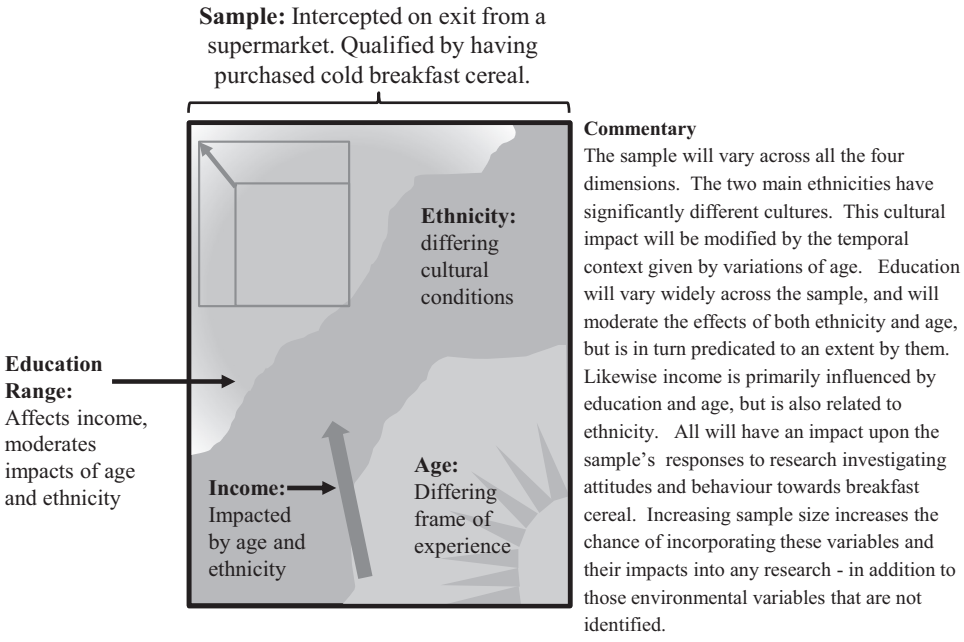

Fig. 7. Transposing the agricultural research environment to social sciences.

obtained to other significantly different environments is required, then replications of the entire exercise must be conducted to establish its validity across that range.

Thus while the individual experimental units may be small, the number of replications that might be involved means that overall research exercise to test the validity of a single hypothesis, and to then extend it over a meaningful range of environments may be extremely large. Thus it must be stressed that the 'small sample' required by Fisher's principles applies to the individual experimental unit, and not to the overall research exercise that may be required to adequately test any specific hypothesis. It therefore does not equate in any way to a social scientific 'free lunch'.

8. Applying Fisher's principles to marketing: an example

In order to demonstrate the application of Fisher's principles to marketing research, a recent publication in the nutrition literature will be used (Hamlin and McNeill, 2016). This research sought to test whether the recently introduced Australasian Health Star Rating System (HSR) had any impact upon consumer choice. This front of pack (FoP) nutritional information labelling system had been introduced across Australasia at a cost of several million dollars, with no prior consumer testing.

The first task was to refine the research hypotheses; minimising them so that a contribution to knowledge could be made by applying the most efficient experimental design to the smallest possible sample size. The untested introduction of this system implied an assumption on the part of those introducing it that the 0.5 to 5 star rating displayed by the system would have a significant impact upon consumer choice at the point of sale 'ceteris paribus', and that that impact would also be significantly correlated to the number of stars displayed (0.5 stars-negative to 5.0 stars-positive). The lack of a segmentation or targeting strategy for the system indicated that it was also assumed that the system would impact upon all consumers and all products equally.

The presence of the second assumption allowed the researchers to make a general contribution by applying a very narrow test to the first assumption under a single very highly specific set of conditions. The research hypotheses were therefore defined as:

H¹. The HSR FoP label would significantly influence consumer choice.

H². The HSR FoP labels' impact upon consumer choice would be moderated by variations in the 0.5–5 star rating expressed by the label.

And were tested for the single case of high and low nutritional value muesli products. The smallest and simplest experimental unit that could be used to test these hypotheses was identified as a 2 product (high/low nutritional status) × 2 HSR label (present/absent) full factorial with two internal replications in each cell, each containing 50 consumer choice observations (Fig. 8). ***H¹*** would be supported by the presence of a main effect for the label treatments, and ***H²*** would be supported by the presence of a significant interaction of the right type between products and labels. In a 2 × 2 factorial design the nature of any interaction can be uniquely established by visual plotting. The dependent variable was *unprompted* consumer choice, essential for an effective test of the label, but a rarity in this area of research (Lachat and Tseng, 2013).

The use of a common comparator product as one pole of all the choice tasks meant that each consumer recruited could undertake two choice tasks without becoming aware of the purpose of the research and thus prompted by it. This halved the required number

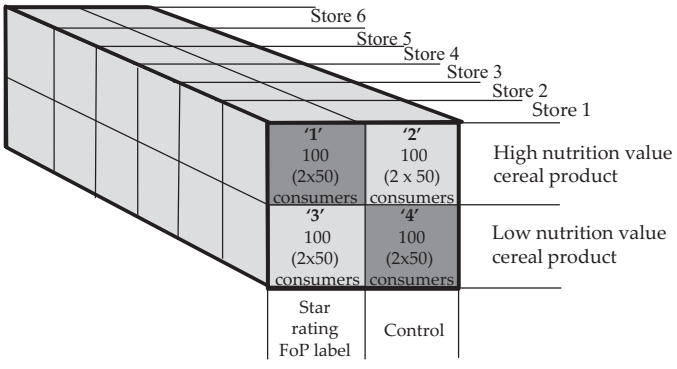

Fig. 8. Experimental design.

Please cite this article in press as: Robert Hamlin, How small sample size and replication can increase accuracy in experiments: Lessons that marketing may learn from agricultural scientific method, Australasian Marketing Journal (2017), doi: 10.1016/j.ausmj.2017.05.002

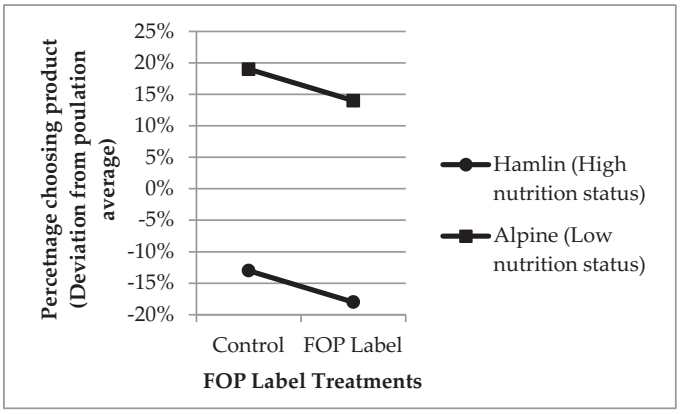

Fig. 9. Experimental result.

of consumers for each experiment from 400 to 200. The figure of 50 consumers per cell is a minimum that has been arrived at by successfully (and unsuccessfully) applying research of exactly this type for nearly 30 years – the minimum figure depends upon the experimental design that is used. The two products were similar cold cereal products that only differed markedly in their nutritional status. The FoP label, when present, merely expressed that status.

The prime factor that was kept stable was personal situation when tested. Thus all consumers were intercepted on exit from a supermarket at which they had purchased cereal. Consumers exiting from retail stores tend to vary considerably, depending upon the time of day and the day of the week. Therefore in order to acquire as homogeneous a sample as possible, recruitment involved an incentive, which when coupled with a short task, allowed a very short window for each sample's recruitment (typically 2–3 hours for a set of 200). The individual 2×2 experimental design was replicated six times in individual field environments that were as similar to each other as could be achieved by the researchers. All samples were recruited starting at 9.00 am the same day of the week, at six similar branded retail stores in the same city in New Zealand. Every available site that matched the researcher's criteria for similarity was utilised.

The outcome of the exercise is summarised in the chart in Fig. 9. The statistical analysis merely confirmed what is visually apparent, namely that there was a significant (negative) main effect for the labels, thus supporting $\boldsymbol{H^1}$, but that the interaction between the high and low nutritional value products and their respective labels was virtually non-existent, thus not supporting $\boldsymbol{H^2}$. This outcome is perhaps not 'rich' in its volume, but it is certainly rich in its implications. The results show that the FoP label was attended to, but that it did not function in the way that its creators intended. The number of stars presented by the FoP label had no effect on consumer choice, thus rendering it effectively useless in this particular environment.

This outcome is, within its very specific limits, highly robust. It has a high degree of internal validity given its stability to internal replication, this is a source of support that would not have been available had a single larger design been used with the same 1200 participant sample. It is thus hard to either challenge or attach any other conclusion to the outcomes of this specific exercise due in no small degree to its tight focus and simplicity. This is important when a result is likely to be contentious.

The external validity (generalisability) of this research is not open to challenge because it makes no claims to that end. However, the very high demonstrated level of internal validity of the initial research result creates a reliable basis for establishing the limits of this finding's applicability by further systematic replication and development of the entire exercise. It is less likely that this solid basis would have been available had the researchers attempted to achieve a 'generalisable' result via a single exercise by introducing wide variations into the replications (different groups/stores/cities) or into the individual experimental units (different products/labels/groups).

9. Small sample sizes, the role of replication, life science and marketing theory

While the concept of quantifiable statistical risk within experiments is well understood within marketing research its unavoidable companion in life sciences, unquantifiable experimental environmental uncertainty, is not. Environmental uncertainty can only be reduced by care in design and execution and minimum sample sizes. As techniques based upon parallel comparison represent a significant proportion of all published research in the marketing discipline, the idea that the optimal sample size for an experimental exercise may not be the largest that resources allow, but is in fact the smallest that the most efficient that a tight hypothesis and an appropriate methodology permits, is worthy of further dissemination.

A related observation is that if circumstances within research situations involving these techniques do require a larger dataset, then if Fisher's principles are applied to the design, this increased size is usually achieved by intra-study replication of smaller experimental patterns, which allows any environmental irregularities impacting upon the larger design to be partitioned and analysed either as sum of squares for replication, or by treating replication as a true independent variable in the analysis of variance table. It thus allows the stability, or otherwise, of the result within a specific environment to be effectively demonstrated within a single exercise. It is important to appreciate that both of these outcomes have significant value.

While this well-proven practice of intra-study replication is common in agricultural research (e.g. Fig. 6), it is rare in published academic marketing research (Uncles and Kwok, 2013a, 2013b), which means that marketing is deprived of potentially important insights from intra-study replications. The recent article published by Kwon et al. (2017) in the European Journal of Marketing that seeks to dismiss the value of intra-study replications on the basis that they are not conducted by 'independent' researchers is a particular ominous development in this regard, that cannot be reconciled with its routinely successful application in both commercial and academic agricultural science over an extended time period.

The use of small sample sizes in response to environmental variations also has a bearing on inter-study replication. The appreciation of the uncertainty created by such environments led to a different understanding of the role of inter-study replication within the agricultural and related life sciences. It was appreciated that in the face of such variation a truly generalisable result might not be achievable, and even if it was achieved, it might not be truly meaningful. Thus, rather than requiring inter-study replication to confirm universal theoretical 'truths' via outcomes that were fully consistent with the initial research, replication outcomes that were not consistent with the initial work were appreciated to be valuable sources of additional information. They indicated the possible environmental moderation of and limits to the 'local truth' of any initial result, rather than an automatic damning of either the initial result or the replication.

This observation has the capacity to inform the debate within the marketing literature on the very low published rates of replication of both types within the discipline (Hubbard and Armstrong, 1994; Hubbard and Vetter, 1996; Kerr et al., 2016; Lehmann and Bengart, 2016; Morrison et al., 2010; Park et al., 2015). Some of these concerns have been expressed with exceeding bluntness. "*As things*

Please cite this article in press as: Robert Hamlin, How small sample size and replication can increase accuracy in experiments: Lessons that marketing may learn from agricultural scientific method, Australasian Marketing Journal (2017), doi: 10.1016/j.ausmj.2017.05.002

now stand, practitioners should be skeptical about using the results published in marketing journals as hardly any of them have been successfully replicated, teachers should ignore the findings until they receive support via replications and researchers should put little stock in the outcomes of one-shot studies" (Evanschitzky et al., 2007, p. 411).

The quote above, which has 175 citations at the time of writing, inadvertently identifies a key issue with the philosophy of marketing science, namely that, given its working environment, it expects too much from replication before it is considered to be "*successful*". 'Success' in replication does not require consistency of outcome, but that an outcome, whatever its nature, is theoretically informative. One plausible and observable result of this unrealistic expectation is a well-founded aversion to (fear of?) replication of any type within marketing research, due to the unachievable theoretical expectations and potentially highly negative inter-researcher implications that may be applied to some of its possible 'unsuccessful' outcomes. Unachievable peer expectations lead to a failure to publish 'unsuccessful' replication research, however, informative it may be. Failure to publish replication research in the current world of academic performance metrics and elusive tenure rapidly leads to a perfectly reasonable learned aversion to initiating it, either as a stand-alone activity or as an integral part of any research exercise.

Academic marketing can only achieve a degree of integrity by adapting its theoretical expectations to a format that is consistent with research outcomes that are practically achievable within its environment through the use of these experimental tools and by using small samples and routine intra and inter-study replication to investigate the limits of such generalisations as can meaningfully be described as useful 'local truths'. One sign that such an approach is being adopted will be a rise in the number of replications that are reported in the literature. As there is no sign of this occurring at present, with the already very low reported rate of replication actually declining (Evanschitzky et al., 2007), marketing science is likely to continue to struggle to achieve the same level of progress as those life science disciplines, such as agriculture that currently apply a more pragmatic theoretical framework to research undertaken in a similarly uncertain and unpredictable working environment.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Banks, S., 1965. Experimentation in Marketing. McGraw-Hill, Boston.
Berzsenyi, Z., Gyrffy, B., Lap, D., 2000. Effect of crop rotation and fertilisation on maize and wheat yields and yield stability in a long-term experiment. Eur. J. Agron. 13 (2–3), 225–244. doi:10.1016/S1161-0301(00)00076-9.
Brunk, M.E., Federer, W.T., 1953a. How marketing problems of the apple industry were attacked, and the research results applied. In: Methods of Research in Marketing No. 4. Department of Agricultural Economics, Cornell University, Agricultural Experiment Station, New York State College of Agriculture, Ithaca, NY. Pamphlet.
Brunk, M.E., Federer, W.T., 1953b. Experimental designs in probability sampling in marketing research. Am. Stat. Assoc. J. 440–452. doi:10.1080/01621459.1953.10483484.
Calder, B.J., Phillips, L.W., Tybout, A.M., 1981. Designing research for application. J. Consum. Res. 8 (2), 197–207. doi:10.1086/208856.
Cox, K., 1964. The responsiveness of food sales to shelf space changes in supermarkets. J. Mark. Res. 30, 63–67. doi:10.2307/3149924.
Dodds, W.B., Monroe, K.B., Grewal, D., 1991. Effects of price brand and store information on buyers' product evaluations. J. Mark. Res. 28, 307–319. doi:10.2307/3172866.
Easley, R.W., Madden, C.S., 2013. Replication revisited: introduction to the special section on replication in business research. J. Bus. Res. 66 (9), 1375–1376. doi:10.1016/j.jbusres.2012.05.001.
Easley, R.W., Madden, C.S., Dunn, M.G., 2000. Conducting marketing science: the role of replication in the research process. J. Bus. Res. 48 (1), 83–92. doi:10.1016/S0148-2963(98)00079-4.
Easley, R.W., Madden, C.S., Gray, V., 2013. A tale of two cultures: revisiting journal editors' views of replication research. J. Bus. Res. 66 (9), 1457–1459. doi:10.1016/j.jbusres.2012.05.013.
Evanschitzky, H., Baumgarth, C., Hubbard, R., Armstrong, J.S., 2007. Replication research's disturbing trend. J. Bus. Res. 60 (4), 411–415. doi:10.1016/j.jbusres.2006.12.003.
Fisher, R.A., 1925. Statistical Methods for Research Workers. Oliver & Boyd: E, Edinburgh.
Fisher, R.A., 1935. The Design of Experiments. Oliver & Boyd, Edinburgh.
Fisher, R.A., 1950. Contributions to Mathematical Statistics. John Wiley & Sons Ltd, New York.
Gomez, K.A., Gomez, A.A., 1984. Statistical Procedures for Agricultural Research. John Wiley & Sons, New York.
Hamlin, R., McNeill, L., 2016. Does the Australasian "health star rating" front of pack nutritional label system work? Nutrients 8 (6), 327–341. doi:10.3390/nu8060327.
Hamlin, R.P., 1997. The Meat Purchase Secision (Unpublished Ph.D. thesis). University of Otago, Dunedin, NZ.
Hamlin, R.P., 2005. The rise & fall of the Latin square in marketing, a cautionary tale. Eur. J. Mark. 39 (3/4), 328–350. doi:10.1108/03090560510581809.
Hubbard, R., Armstrong, J.S., 1994. Replications and extensions in marketing: rarely published but quite contrary. Int. J. Res. Mark. 11 (3), 233–248. doi:10.1016/0167-8116(94)90003-5.
Hubbard, R., Vetter, D.E., 1996. An empirical comparison of published replication research in accounting, economics, finance, management, and marketing. J. Bus. Res. 35 (2), 153–164. doi:10.1016/0148-2963(95)00084-4.
Jones, D.L., Rousk, J., Edwards-Jones, G., DeLuca, T.H., Murphy, D.V., 2012. Biochar-mediated changes in soil quality and plant growth in a three year field trial. Soil Biol. Biochem. 45, 113–124. doi:10.1016/j.soilbio.2011.10.012.
Kennedy, J.R., 1970. The effect of display location on the sales and pilferage of cigarettes. J. Mark. Res. 7, 210–215. doi:10.2307/3150111.
Kerr, G., Schultz, D.E., Lings, I., 2016. Someone should do something: replication and an agenda for collective action. J. Advert. 45 (1), 4–12. doi:10.2307/3150111.
Koschate-Fischer, N., Schandelmeier, S., 2014. A guideline for designing experimental studies in marketing research and a critical discussion of selected problem areas. J. Bus. Econ. 84 (6), 793–826. doi:10.1007/s11573-014-0736-2.
Kwon, E.S., Shan, Y., Lee, J.S., Reid, L.N., 2017. Inter-study and intra-study replications in leading marketing journals: a longitudinal analysis. Eur. J. Mark. 51 (1), 257–278. doi:10.1108/EJM-07-2015-0450.
Lachat, C., Tseng, M., 2013. A wake-up call for nutrition labelling. Public Health Nutr. 16 (3), 381–382. doi:10.1017/S1368980012005666.
Langton, S., 1990. Avoiding edge effects in agroforestry experiments; the use of neighbour-balanced designs and guard areas. Agroforestry Syst. 12 (2), 173–185. doi:10.1007/BF00123472.
Lehmann, S., Bengart, P., 2016. Replications hardly possible: reporting practice in top-tier marketing journals. J. Model. Manag. 11 (2), 1–28. doi:10.1108/JM2-04-2014-0030. (Preprint).
Montaguti, E., Neslin, S.A., Valentini, S., 2015. Can marketing campaigns induce multichannel buying and more profitable customers? A field experiment. Mark. Sci. 35 (2), 201–217. doi:10.1287/mksc.2015.0923.
Morrison, R., Matuszek, T., Self, D., 2010. Preparing a replication or update study in the business disciplines. Eur. J. Sci. Res. 47 (2), 278–287. DOI: N/A.
Olson, J.G., McFerran, B., Morales, A.C., Dahl, D.W., 2016. Wealth and welfare: divergent moral reactions to ethical consumer choices. J. Consum. Res. 42 (6), 879–896. doi:10.1093/jcr/ucv096.
Orth, U.R., Malkewitz, K., 2012. The accuracy of design-based judgments: a constructivist approach. J. Retail. 88 (3), 421–436. doi:10.1016/j.jretai.2011.11.004.
Park, J.H., Venger, O., Park, D.Y., Reid, L.N., 2015. Replication in advertising research, 1980–2012: a longitudinal analysis of leading advertising journals. J. Curr. Issues Res. Advert. 36 (2), 115–135. doi:10.1080/10641734.2015.1023874.
Rui, J., Meyers-Levy, J., 2009. The influence of self-view on context effects: how display fixtures can affect product evaluations. J. Mark. Res. 46 (1), 37–45. doi:10.1509/jmkr.46.1.37.
Sunding, D., Zilberman, D., 2001. The agricultural innovation process: research and technology adoption in a changing agricultural sector. Handb. Agric. Econ. 1, 207–261. doi:10.1016/S1574-0072(01)10007-1.
Uncles, M.D., Kwok, S., 2013a. Designing research with in-built differentiated replication. J. Bus. Res. 66 (9), 1398–1405. http://dx.doi.org/10.1016/j.jbusres.2012.05.005.
Uncles, M.D., Kwok, S., 2013b. Reply to commentary on designing research with in-built differentiated replication. J. Bus. Res. 66 (9), 1409–1410. doi:10.1016/j.jbusres.2012.05.007.

Robert Hamlin is a graduate of the University of Oxford, where he studied agricultural science. He is currently a senior lecturer at the Department of Marketing, University of Otago. Over the last 20 years he has studied low involvement consumer decision making, and consumer response to food package design and point of sale communications. His specific focus is on the use of experimentation to study revealed consumer behaviour. He has a particular interest in the acquisition of very clean data by applying best experimental practice in the field. Prior to retraining in business studies at Indiana University while holding a Harkness Fellowship, he held management/research positions at the Pig Improvement Company, and the University of Oxford Agricultural Field Research Station.

Please cite this article in press as: Robert Hamlin, How small sample size and replication can increase accuracy in experiments: Lessons that marketing may learn from agricultural scientific method, Australasian Marketing Journal (2017), doi: 10.1016/j.ausmj.2017.05.002