Experimental design and statistical analysis for on-farm research with resource-poor farmers: a practical review

Brian E. Love
Laki Goonewardene
Dean Spaner

ABSTRACT

The Green Revolution has greatly increased agricultural production but has neglected resource-poor farmers. Farming systems research and extension (FSRE) research approaches have been adopted to address this concern. On-farm research consisting of: 1) preliminary surveys and 2) on-farm trials, is a central component of FSRE but often requires untraditional experimental design and statistical analyses. This practical review outlines experimental design considerations and statistical techniques for on-farm research.
Introduction

Green revolution

The Green Revolution (1960s) produced high yielding crop varieties (Clawson and Hoy, 1979), but has neglected certain regions and farmers (Evenson and Gollin, 2003) and is dependent on the use of agrochemical inputs (fertilizers, pesticides, etc.). Generally, resource-poor farmers practicing rainfed agriculture have not benefited (Greenland, 1975). Poor adoption has been related to risk-adverse behaviour due to marginal economic situations (Clawson and Hoy, 1979) and uncertain farming environments (Simmonds, 1986). This is a concern because globally 450 million resource-poor farmers support 1.25 billion people (Mazoyer, 2001).

Farming systems research and extension

Farming systems research and extension (FSRE) approaches developed as a response to evidence, which emerged in the 1970s, that poor farmers were not adopting Green Revolution technology (Simmonds, 1986). The key components of this approach are: 1) multidisciplinary diagnosis of farmers’ practices and problems, 2) laboratory and on-station research, 3) on-farm biophysical research, 4) socio-economic research and evaluation of on-farm trials and the farming community, 5) rapid dissemination and diffusion of results (Stroup et al., 1993). The effectiveness of FSRE was questioned in the late 1980s (Fielding, 1988; Herdt, 1987) and structural adjustment policies have reduced its priority (Finan, 1993), but it has remained a successful approach for working with poor farmers when properly applied (Tripp, 1991).
On-farm research is a fundamental component of FSRE (Riley and Alexander, 1997; Simmonds, 1986; Stroup et al., 1993; Tripp, 1991) and facilitates technology adoption (Eremie et al., 1991; Pillay, 1999). In developing countries where research station conditions differ greatly from farm conditions on-farm research is required to produce credible results (Parkhurst and Francis, 1986). On-farm research consists of preliminary studies of existing farming systems and subsequent on-farm trials to test research ideas (Mutzaers and Walker, 1991; Simmonds, 1986). On-farm experiments often use untraditional experimental designs and their interpretation lies beyond commonly taught statistical techniques (Riley and Alexander, 1997). On-farm experiments involve problems not encountered in on-station trials (Fielding and Riley, 1998), contain greater variability (Fielding and Riley, 1998; Riley, 2000), and generate a combination of qualitative and quantitative data (Riley and Alexander, 1997).

Past reviews

A review of statistics in international agriculture found statistical standards to be poor (Lauckner, 1989). Riley and Alexander’s (1997) review of on-farm trial design and statistical analysis found that most sources (e.g. Gomez and Gomez, 1984; Hammerton and Lauckner, 1984; Hildebrand and Poey, 1985; Neeley et al., 1991; Steiner, 1987) presented traditional designs typically used in on-station trials, with only Mutzaers and Walker (1990) and Stroup et al. (1993) providing useful and flexible approaches. Their review was literary rather than
practical in nature. Eberhardt and Thomas’s (1991) review of ecological field studies and
Chibnik’s (1985) review of statistics in sociocultural anthropology can also be applied to on-
farm work.

Research is either mensurative (sampling existing phenomenon) or manipulative
(applying treatments) (Hurlbert, 1984). On-farm mensurative experiments may be used to
develop hypotheses for manipulative experiments (Settle et al., 1996). In this review, surveys
(preliminary studies) are considered synonymous with mensurative experiments while trials are
considered synonymous with manipulative experiments.

Preliminary study of farming systems: experimental design

History

It is important to understand existing farming systems (Simmonds, 1986). Traditionally
formal survey methods (e.g. Poate and Daplyn, 1993) have been employed. Recently, new
methods have been developed. Rapid rural appraisal (RRA) (Carruthers and Chambers, 1981)
came about in part as a response the ineffectiveness of costly large-scale surveys that often
produced misleading, difficult to use, and largely ignored results (Chambers, 1994a). RRA has
subsequently evolved into participatory rural appraisal (PRA) because it was found that rapid
techniques were not inclusive (Mascarenhas, 1991). RRA and PRA approaches shifted survey
techniques from: outsider (researcher) to insider (farmer) explanations, measuring to comparing,
closed to open questionnaires, individual to group interviews, and verbal to visual interactions
(Chambers, 1994b).
The origins (Chambers, 1994a), experiences (Chambers, 1994b), and challenges and potential (Chambers, 1994c) of participatory rural appraisal have been reviewed. Few manuals on methods exist due to the perception that common sense should guide the researcher (Chambers, 1994a), but Chambers (1994a; 1994b) and Mascarenhas (1991) outline common PRA techniques. Despite these shifts formal survey approaches remain important.

Experimental design

Surveys can be either descriptive or comparative (Oppenheim, 1992). Data from descriptive surveys can be used post-hoc to conduct comparisons. Issues of sampling and measurement apply across all survey types.

Sampling requires definition of the sampling universe and its sampling units (population). In agriculture, sampling units can be farmers, fields, etc. with the sampling universe being all sampling units of interest based on specific attributes (e.g. receiving credit, cooperative member, etc.) (Poate and Daplyn, 1993). Sample frames (database of sampling units) are required for sample selection. Agricultural surveys have used both area frames (agricultural land) and list frames (known farmers) (Yates, 1960). List frames tend to be incomplete but more efficient than area frames for sampling farms (Chhikara and Lih-Yuan, 1992). Sample frames are often incomplete (Cochran, 1977) and/or inaccurate (Poate and Daplyn, 1993).

Sample size is influenced by whether a survey is descriptive or comparative. Descriptive studies aim to provide accurate estimates of population parameters. Required sample size depends on: 1) heterogeneity of the population, 2) number of factors (strata) studied, 3) size of the investigated parameter, and 4) the desired level of precision for the estimate (Bernard,

When comparisons rather than estimates are required sample sizes are determined by: 1) the statistical test being used, 2) amount of variability in the compared populations, 3) the size of the difference to be detected, 4) and the desired probabilities of detecting true differences and false differences (Sokal and Rohlf, 1995). When comparing proportions variability estimates are not required. Zar (1999) provides mathematical formulas for sample size calculations sizes for many statistical tests.

The information required for sample size calculations is based on the objectives of the experiment (number of strata) and the judgment of the researcher (desired: precision, probabilities, size of the detectable difference, and the test used). Variability and parameter size estimates may be based on researchers’ best guesses (Bernard, 2002), past studies (Gomez and Gomez, 1984), and/or pilot studies (Cochran, 1977).

Sampling techniques

There are many techniques for sampling populations. Simple random sampling, a form of probability sampling in which each sampling unit has an equal probability of being selected) permits the generalization of results to the study population (Poate and Daplyn, 1993). Prior knowledge of population attributes permits sample stratification (division of sampling units into groups/strata), which can improve efficiency (Yates, 1960). Sampling within strata may be proportional to the stratas’ sizes, equal for all strata, or unequal (Poate and Daplyn, 1993; Yates,

Simple random sampling and stratified random sampling assume sample frame availability, which is often not the case. Multi-stage sampling or cluster sampling may be employed using alternative frames (e.g. towns) and appropriate frames developed after a few levels of sampling by taking a census (Poate and Daplyn, 1993; Bernard, 2002). Such sample frame construction through censuses is costly, but up to a third of the sample size should be sacrificed to construct sample frames for large-scale surveys (Scott, 1985).

Alternatively, smaller and smaller clusters may be sampled until the sampling unit of interest is obtained (Bernard, 2002). For instance towns, then streets, and then houses may be selected to obtain a sample of farmers. When sampling fields random line intercepts based on compass bearings may be used (Yates, 1960). Sampling units of the same cluster (country, district, town, etc.) are more alike (Poate and Daplyn, 1993). Thus, sampling a few units (e.g. farmers) from each of a large number of clusters (e.g. towns) is preferred (Bernard, 2002).

Rural appraisal techniques often use purposive instead of probability sampling procedures (Carruthers and Chambers, 1981; Mascarenhas, 1991). This is because the studies seek to examine a number of factors (strata) and the populations under study are often scattered and lack sample frames (Carruthers and Chambers, 1981). Bernard (2002) describes a number of nonproportional sampling techniques. He notes that while they produce biased population estimates they are less costly and can be useful for: 1) gaining information on issues to be studied (convenience sampling), 2) studying social networks and building sampling frames.
(snowball sampling), 3) providing a similar control group (case control sampling), and 4) collecting cultural data (key informant sampling).

Key informants must be competent. Poate and Daplyn (1993) define informants as being observers, authorities, or doers and suggest that response accuracy increases in that order. Romney et al. (1986) developed a cultural consensus model to test potential informants’ competency in cultural domains. Similarity matrices are constructed based on responses to a test and factor analysis (study of order and structure in multivariate data (Kline, 1994)) is conducted to produce competency scores.

Farmers in developing countries are often very heterogenous (Crossa et al., 2002) as are their environments (Hildebrand, 1984). Defining relatively homogenous groups (recommendation domains, target areas) facilitates the extrapolation of results but there is little consensus on how to define recommendation domains (Franzel, 1992). Recommendation domains may be constructed using multivariate empirical procedures (Crossa et al., 2002), non-local experts (agronomists, scientists, etc.) (Wotowiec et al., 1987), or local experts (farmers) (Ashby, 1990). Use of local experts is advantageous because it is low cost, quick, and permits a relationship to be established with farmers (Franzel, 1992). Recommendation domains have been largely defined by environmental variables (Gomez and Gomez, 1984). More recently socio-economic variables (Crossa et al., 2002) and management practices (Wotowiec et al., 1987) are being used. Defining recommendation domains permits appropriate stratification of samples for both probability and purposive sampling techniques.

Surveying people: interviews
Interviews are used to gather information from people and may be unstructured, semi-structured or structured in format (Bernard, 2002). Unstructured interviews have a topic of interest but otherwise exert minimal influence on informants in order to facilitate openness and expression. Semi-structured interviews are guided by a list of questions/topics that need to be covered but allow interviewers to follow leads and probe. Structured interviews attempt to get informants to respond to a uniform set of stimuli (lists of questions in a particular order). Rural appraisal methods prefer guided and group interviews to structured interviews (Carruthers and Chambers, 1981), while traditional agricultural surveys prefer structured questionnaires (Poate and Daplyn, 1993). Bernard (2002) provides practical and technical advice on how to conduct different types of interviews.

Questions and the order in which they are asked (organization) are the fundamental components of questionnaires (structured interviews). Good questions are simple, understandable, bias-free, and non-irritating (Payne, 1973). Questions may be open-ended or closed with open-ended questions either permitting any response or selection from a list of predetermined responses (Labaw, 1980). Bernard (2002) provides numerous suggestions for question wording including that they be unambiguous, use appropriate vocabulary, have a clear purpose, provide appropriate alternatives, are short when unthreatening, and are not leading. Threatening questions result in underreporting (Sudman and Bradburn, 1974) but response estimates can be improved by assessing threat perception (Bradburn et al., 1978). Questions regarding land title, livestock possession, etc. may be threatening in certain contexts.

Different question wording can result in differences of up to 19 percentage points with improvements in question wording contributing far more to accuracy than improvements in sampling methods (Payne, 1973). Labaw (1980) argues that questionnaires aimed at predicting
behavior (e.g. technology adoption) should focus on questions about informants’ environment, knowledge, and actual behavior because attitudinal questions reveal little about future behavior.

Questionnaire format is also important. The format of a questionnaire should: 1) prevent position bias, 2) provide smooth flow, 3) be easy to follow for interviewer, and 4) be efficient for data entry (Labaw, 1980). Order of questions can affect the responses given in surveys (Noelle-Neumann, 1970; Schuman et al., 1983). Use of buffer questions (relatively neutral questions) preceding sensitive questions can improve the truthfulness of responses (Thumin, 1962). Neutral questions have also been used to mitigate context effects in which a preceding question alters the response to the next question although sometimes context effects are impossible to mitigate (Schuman et al., 1983). Grouping like questions can affect responses but Metzner and Mann (1953) found this not to be the case and suggest that randomly dispersing questions can frustrate informants. Overall, balance is required when designing surveys with regards to length, repetition, and ordering of questions.

Ranking or scoring, are often used for opinion questions (Bernard, 2002). Ranks place items in order while scoring assigns points to items. Farmer preferences can be evaluated with these techniques (e.g. Abeyasekera et al., 2002; Ashby, 1990; Ashby et al., 1987; Bellon, 1996; del Pilar Guerrero et al., 1993). Ranking does not provide information on the magnitude of the gaps between ranks, which negates evaluation of relative preference (Maxwell and Bart, 1995).

Preference evaluation techniques should permit ties, extreme values, and be easy to administer (Fielding and Riley, 2000). Ranking fails the first two criteria, but scoring meets all three requirements when five times as many points as items are used (Fielding and Riley, 2000). Scoring can be time consuming, for instance 75 minutes to score ten traits for four items.
(Abeyasekera et al., 2002). Maxwell and Bart (1995) describe a number of different types of
scoring. Spaner (1997) used a mark drawn on a line in order to produce scores on a continuous
scale. Ranking can be facilitated by sorting cards (Ashby, 1990). Pair-wise ranking in which all
pairs of items are ranked one at a time may make ranking easier for farmers (Fielding et al.,

Pre-testing surveys on a small-scale usually results in improvements and can reveal
problems before they become unmanageable at larger scales (Cochran, 1977). Pre-tests are
conducted to evaluate individual questions and ensure that they form a consistent well organized
questionnaire in the case of structured interviews (Oksenberg et al., 1991). Pre-testing should be
conducted under conditions identical to those of the survey and pre-test informants must not be
used in the survey (Bernard, 2002). Even after pre-testing it is often not possible to agree upon
appropriate modifications (Suchman and Guttman, 1947) and pretests can be too small to detect
important errors (Payne, 1973). Regardless pre-testing is important and Oksenberg et al. (1991)
review its rationale, implementation, and recent improvements.

Interview context (physical/social setting) influences informants’ responses. In on-farm
research deception during interviews can be a problem (Malton, 1983). Prepaid non-monetary
incentives can increase response rates and improve response completeness (Willimack et al.,
1995). Active involvement of informants (e.g. card sorting) reduces monotony and can result in
improved responses (Noelle-Neumann, 1970). If the interviewer is of a different ethnic group,
responses may be affected especially for informants of low-income status (e.g. poor farmers)
(Schuman and Converse, 1971). Group discussion in group interviews (focus groups) can inhibit
individuals and result in false consensus but is productive for exploratory and feedback work
Interviewing informants after events occur can result in forgetting events or telescoping of events in time. Telescoping is the placing of an event forwards or backwards in time in comparison to its actual time of occurrence (Neter and Waksberg, 1964). Analysis of aggregate telescoping demonstrates a forward effect (Neter and Waksberg, 1964; Sudman et al., 1984), that is events are recalled as occurring more recently than is actually the case. Recall can be improved by bounding the recall period. For multiple interviews the visit schedule bounds recall while for one time interviews boundaries must be verbally delimited before each set of questions. Longer recall periods are more resource efficient but recall periods of greater than 1 month result in events being forgotten (Neter and Waksberg, 1964).

If the event is fixed in the farmer’s mind through participation recall can be improved (Spencer, 1991). Visits must be appropriately timed to facilitate recall (planting, harvesting, etc.) (Ashby, 1990), 2 to 4 well-timed visits are thought to be adequate (Byerlee and Triomphe, 1991; Versteeg and Huijsman, 1991), but weekly (Hildebrand, 1984) and bi-monthly (Cobbina and Atta-Krah, 1991) visits have also been advocated. Timely visits are difficult to make and extra visits are usually required to ensure contact (Stroud, 1993). Regular contact may inconvenience participants and assuring participation may become difficult (Oppenheim, 1992). Resource availability and the cost-benefit analysis of more accurate information weighed against increased costs determine visit frequency.

Surveys may sample objects (e.g. crop fields, soil, etc.) rather than people. After sample unit selection sub-samples are commonly taken (Poate and Casley, 1985). Systematic sampling for sub-samples can give better precision than random sampling (Rao and Coe, 1991). Quadrats
are often used to take samples and the number of quadrats and their size depends on variability. Completely random (uncorrelated) variability is as precisely measured with a single large quadrat as with many small quadrats and reduces costs, and measurement error (Rao and Coe, 1991). Most biological phenomenon are in some way spatial structured (non-random) (Dale, 1999) making multiple small quadrats preferable. Precise yield estimates require at least three sub-samples but generally coefficients of variation remain high ~50% (Poate and Casley, 1985). If correlation between subsamples is greater than 0.7 and/or the coefficient of variation is <40%, subsamples likely lack independence (Poate and Casley, 1985). Increasing distance between quadrats remedies this. For diversity estimates many small quadrats accurately estimate abundance for common species but yield incomplete species lists whereas a few large quadrats result in completer species lists but overestimate the abundance of rare species and give imprecise abundance estimates for common species (McCune and Lesica, 1992).

Quadrats may be of various shapes. Circular plots have the lowest area to circumference ratio reducing the number of boarder inclusion judgments required and can be marked permanently by a single point (McCune and Grace, 2002), but are often difficult to place (Poate and Casley, 1985). In on-farm agroforestry trials, quadrats tend to be rectangular but may be row lengths as well (Rao and Coe, 1991). Positioning a rectangular quadrat over a row-planted crop requires that the corners of the quadrat fall along the one of the rows to prevent row exclusion (Poate and Casley, 1985). Other sampling units such as point intercept, line intercept and distance methods are alternatives to quadrat sampling (McCune and Grace, 2002).

Remote sensing techniques are cost-effective and non-destructive (Gerard and Buerkert, 2001). Remote sensing may be carried out by terrestrial fixed platforms (e.g. moveable booms), aircraft, and satellites (Shanahan et al., 2001). Tucker (1979) reviews the development of remote
sensing techniques for biomass estimation. Lawerence et al. (2000) developed a bicycle mounted radiometer for estimating biomass in plots. Alternatively, satellite information may be used when larger areas are involved (Idso et al., 1977). Forecasting yield for regions requires both area and production estimates (Tennakoon et al., 1992). Visual estimation of crop output is also non-destructive (CIMMYT, 1984) and does not involve expensive equipment.

An alternative to physically sampling production is farmer reporting, which gives results comparable to crop cutting methods and is more practical (Poate and Daplyn, 1993). Farmers and professionals alike find it difficult to estimate crop area (Poate and Casley, 1985). Thus crop-cutting may be the preferred for yield estimates because area is controlled precisely by quadrat size. Geographical positioning system units (GPS units) provide an alternative for on-farm area estimation.

On-farm trials: experimental design

In manipulative experiments experimental designs are defined by how they organize replicates (completely random or random but structured) and by the number of factors (one or many) they attempt to assess. Traditionally agricultural experimental designs have attempted to control for local variation through the structured organization of replicates (blocking). Ordering of experimental units constitutes the design structure of the experiment. If more than one factor is being assessed the experiment is said to be a factorial experiment. These factors are imposed treatments and may have a number of levels. The number of treatment types and number of levels of treatments within each type are the treatment structure of the experiment. Experimental design is discussed with reference to its components: randomization, treatments, blocking,
replication, etc. Gomez and Gomez (1984) and Pearce (1986) provide reviews of standard
experimental designs for agricultural research.

Numerous typologies based on level of farmer participation, state of knowledge, and type
of technology have been proposed for classifying on-farm trials (Mutsaers et al., 1991). Riley
and Alexander (1997) identify three levels of farmer participation regarding research; 1)
researcher planned and implemented with farmers providing preference information and land, 2)
researcher planned and farmer implemented, 3) farmer planned and implemented with
researchers observing. In practice there is a continuum of farmer and researcher participation.
Pinney (1991) argues that on-farm trials should evolve towards farmers eventually planning and
implementing experiments after researchers have demonstrated how. Lightfoot (1984) advocates
superimposing treatments on farmers fields. Trial practical aspects vary with farmer
participation, which is important because statistically optimal designs are no good if they are

Farmer involvement requires that treatment number be kept to a minimum (Gomez and
Gomez, 1984; Mutsaers et al., 1991). Mutsaers et al. (1991) suggest that there should be no
more than six treatments. Factorial experiments best address farmers’ questions but require a
large number of treatments (Lafond and Stevenson, <2004>). Various designs address these
conflicting demands. Walker (1991) suggested use of step-wise designs in which treatments are
the sequential addition of factors to a control in order of importance. Unfortunately, step-wise
designs produce meaningless results if the order of factors’ importance is misjudged (Mutsaers et
al., 1991). Gomez and Gomez (1984) proposed a using multiple sets of designs. These sets
include: 1) full factorial design (all combinations), 2) partial factorial design (all factors at the
farmer’s level, all factors at the optimal level, and each factor at the optimal level with all other
factors at the farmer’s level), and 3) two treatment design (all factors at the farmer’s level and all
factors at the optimal level). The ratio of farms receiving each type of design is 1:1:3. Snapp
(2002) advocates mother-baby trials where the full suite of treatments are administered by
technicians (mother) and farmers administer a subset of preferred treatments and a control
(baby).

Treatments should be randomly assigned to plots within farms (Ashby, 1990). However,
farmers may wish to locate certain treatments where they feel they will perform best. When
farmers are involved in the selection of treatments (e.g. Pillay, 1999; Snapp, 2002) treatment
assignment to each farm will be non-random. Grouping of treatment plots into blocks is used to
control for field variation (Pearce, 1986). Farm, in on-farm trials, is analogous to block when
used to group treatments (McIntire and Fussell, 1989). Accordingly, the design is a randomized
complete block or randomized incomplete block if all or a subset of treatments is implemented
on each farm, respectively (Snapp, 2002). If particular farm attributes result in farms being
assigned to treatment classes, e.g. small and large farms, then farm is no longer a blocking factor
because it is a treatment. Replicate plots within farms may be arranged using blocking.

Blocking will only be effective if blocks correspond to underlying environmental
variation (Potvin, 1993). Rainfed agriculture has high levels of within-field variability due to
micro-topography, soil heterogeneity, etc. (Byerlee and Triomphe, 1991). Thus, blocking can be
of questionable value (Mutsaers et al., 1991; Pearce, 1986). In contrast, Fielding and Riley’s
(1998) empirical evaluation of on-farm fertilizer trials in Jamaica found that blocking (especially
lattice designs) within farms increased efficiency and that blocking should be based on farmer
consultation. Contiguous blocks are desirable but may not be possible given the configuration of
farmers’ fields (Gomez and Gomez, 1984). Randomized complete block design efficiency can
be evaluated by quantifying experimental error reduction relative to the experiment analyzed as a
completely randomized design (Steel et al., 1997).

In on-farm trials replicates come in the form of farms and plots within farms. Farm
replicates address among-farm variation while plot replication within farms addresses within
farm variation (Gomez and Gomez, 1984). Desirable amounts of farm and plot replication is
debated. Mutsaers et al. (1991) and Littell et al. (1996) find 20 and 21 farm replicates to be
realistic, respectively. Gomez and Gomez (1984) suggest that the number of replicate farms
should be 10 times the number of cropping systems in the area. Very accurate yield estimates
(confidence intervals of 10% or less of the mean) require 100 replicate farms (Poate and Casley,
1985). Managing 15-25 on-farm trials requires ~2 researchers (CIMMYT, 1982). In practice
many studies have ten or fewer farm replicates (e.g. Pandey et al., 2000; Tornquist et al., 1999).

Commonly there are two replicate plots per farm (Gomez and Gomez, 1984; Mutsaers
and Walker, 1991). Fielding and Riley (1998) found that if 15 or more farms were involved little
was gained from within farm replication. Typically maximizing farm replicates and minimizing
within farm replication is advocated (Mutsaers et al., 1991). Without within farm replication
intra-farm analyses cannot be tested because there is no within-farm error estimate (Cady, 1991).

Stratified random sampling for farm selection is advocated to ensure representativeness
(Gomez and Gomez, 1984; Riley, 2000). Farmer willingness to participate constrains random
selection (Fielding and Riley, 1998; Smith et al., 1991). The effects of non-random selection on
extrapolation of results are not known (Cady, 1991). Typically there is upwards bias to select
larger and more enthusiastic farmers (Riley, 2000). Smith et al. (1991) delimits three levels of
representativeness: farmer, field and management. The importance of each type of
representativeness depends on trial objectives and type of farmer participation. In researcher-
managed trials field representativeness is important but farmer and management
represenativeness are not. Trial farms’ representativeness can be assessed by comparing their
characteristics to those of the farm population they are drawn from (Smith et al., 1991).

On-farm trial plot sizes and shapes vary with trial type. Small plots over-estimate yields
(Riley, 2000). Steiner (1987) recommended 30 m\(^2\) plots for on-farm cereal variety trials.
Fielding and Riley (1998) advocate plot sizes ranging from 75-300 m\(^2\) depending on the crop.
Plots of 1000 m\(^2\) are required if economic analyses (e.g. labor inputs) are to be performed
(McIntire and Fussell, 1989; Spencer, 1991). Stroud (1993) provides a number of plot size
suggestions for different trial types. While larger plots are the norm excessively large plots may
distort farming activities and on-farm trials should not exceed 20% of the farmer’s land (Stroud,
1993). Long narrow plots reduce the variance of yield estimates (Bormann, 1953; Clapham,
1932; Hasel, 1938; Justesen, 1932; Kalamkar, 1932; Pechanec and Stewart, 1940; Wuest et al.,
1994) provided the length of the plot runs in the direction of environmental variation (Hasel,
1938). Irregularly shaped plots may be required to accommodate the shape of farmers’ fields
(Gomez and Gomez, 1984). The strip-plot design consists of vertical and horizontal strips to
which different treatment levels of factors are assigned. This design facilitates farmer
interpretation (Versteeg and Huijsman, 1991) and is suited for two-factor experiments in which
the interaction effect is the priority (Gomez and Gomez, 1984).

Subsampling of plots may be carried out instead of complete harvest. Rao and Coe
(1991) theoretically demonstrate that when comparing means if within-plot variation is small
compared to among-plot variation little is gained by taking more than 5 sub-samples.
Experiments involving trees (agroforestry trials) are an interesting case. Single-tree plots may be
used as replicates (Huxley, 1987) which, negates sub-sampling. For multi-tree plots, Piotto et al.
(2003) based on Wright (1964) advocate sub-samples of 15 trees. Optimizing sub-sampling
designs requires information on metric variability over space and time (Rao and Coe, 1991). If a
factor is contained within another (quadrat samples from plot replicates) the experiment is said to
have a nested design and sub-samples should not be confused with replicates (Hurlbert, 1984).
Nesting may occur at multiple levels (seeds on panicle, panicle in a quadrant, quadrant in a plot).

Covariates are variables that are not explicitly part of the experimental design that are
measured because they may also affect the dependent variable (Lauckner, 1989). In on-farm
research covariates should be used to account for variation (Pinney, 1991) and assess interactions
(Mutsaers et al., 1991). In rotational experiments plots are useful covariates because yield is
correlated on the basis of plot (Singh et al., 1997). When measuring covariates simple scoring
approaches for variables (e.g. shade, weediness) are usually sufficient (Mutsaers, 1991).
Covariates may be measured either at the plot or farm level and Mutsaers (1991) provides
suggestions for appropriate farm and plot level covariates.

Crop production within fields is known to be patchy (Pearce, 1986). Understanding how
the spatial variation of crop production relates to field factors (e.g. soil chemical/physical
properties) is important (Cassel et al., 2000). Crop uniformity trials can be conducted to assess
patchiness (Gomez and Gomez, 1984). Recording spatial information through mapping can be
useful for later interpretation of experiments. In spatial sampling the sampling unit should be
larger than the object of measurement but smaller than uniform patches and the sampling interval
should be smaller than the average distance between patches (Legendre and Legendre, 1998).

Edge effects are a special case of spatial effects. Edge effects occur when the border of
an experimental plot behaves differently (e.g. growth) than its center. Guard rows may be
planted for each plot and at the ends of each block to control for border effects (Durban et al.,
2001). For treatment comparison borders should be removed but for yield estimates borders are important (Rao and Coe, 1991). Border effects can be disproportional in small experimental plots (Langton, 1990) and may distort yield estimates (Clarke et al., 1998).

When edge effects are the result of the neighboring treatment they are known as neighbor effects, which are more difficult to identify than border effects (Langton, 1990). Neighbor effects may be from a spatially or temporally adjacent plot (Bailey and Druilhet, 2004). David et al. (1996) present designs for controlling neighbor effects by grouping treatment plots based on similar characteristics (e.g. plant height). Alternatively, neighbor-balanced designs may be used (Azais et al., 1993; Bailey and Druilhet, 2004).

Statistical analysis

Applied statistics are methods for data collection and analysis, while theoretical statistics provide the framework for understanding the properties and scope of applied methods (Davison, 2003). Applied statistics may be descriptive or analytical. Survey data and experimental trial data analysis share similar underlying principles and are the basis of this review.

Descriptive statistics for surveys provide parameter estimates for populations. These parameters include measures of central tendency (mean, median, mode) and measures of dispersion (range, standard deviation, variance) (Zar, 1999). Parameters are calculated from samples rather than censuses and as such parameter estimates will differ by random chance from the true population value (Bernard, 2002). This uncertainty is referred to as sampling error. Sampling error is estimated by calculating the estimate’s precision as measured by the standard error (Poate and Daplyn, 1993). Confidence intervals can be computed based on the standard error.
error and provide limits within which the true population parameter is known to lie at a particular level of confidence. A number of texts provide formulas for calculating parameter estimates from survey data (Cochran, 1977; Poate and Daplyn, 1993; Yates, 1960). Accuracy is a measure of how close a sample parameter is to the true population parameter (Oppenheim, 1992). Very precise estimates may not be very accurate if they are biased by non-sampling errors. In surveys non-sampling errors may result due to inaccurate sample frames, poor interview design, etc. (Bernard, 2002).

Estimation when comparing one or more groups is very similar to the calculation of population parameters using descriptive statistics. However, when constructing confidence limit estimates, the pooled variance of the groups is used as opposed to the variance of the each individual group (Zar, 1999). Computational intensive resampling techniques (e.g. bootstrap) can be used to generate empirical confidence intervals and do not make assumptions that estimates based on the standard error do (Efron and Tibshirani, 1991).

Analytical statistics make comparisons between populations. These techniques are appropriate for comparative surveys and experiments (trials). Analytical statistics aim to: 1) estimate parameters (discussed above) and 2) test hypotheses (Pearce, 1988). There are many different statistical techniques available to researchers for testing hypothesis. Objectives often determine which techniques are appropriate (Riley, 2000) but statistical techniques are also intimately linked to design. This review focuses on the fundamentals of statistical techniques: hypothesis testing, underlying assumptions, etc. but also addresses important statistical techniques (analysis of variance, non-parametric tests, mixed models, etc.).

Null hypotheses form a baseline for hypothesis testing and are tested against the experimental data and potentially rejected. Probability values (p-values) are calculated from the
experimental data and indicate the probability that the observed event has occurred by chance alone given the null hypothesis (Zar, 1999). To assess the significance of p-values researchers choose a value (α - value) at which p-values will be deemed to indicate significant differences (Sokal and Rohlf, 1995). In hypothesis testing researchers risk committing two types of error: 1) rejecting the null hypothesis when it is true and 2) failing to reject the null hypothesis when it is false (Neyman and Pearson, 1928), which are termed type I (α) and type II (β) error respectively (Sokal and Rohlf, 1995; Zar, 1999). A third type of error (type III error - γ) occurs when a false null hypothesis is rejected but in the wrong direction (MacDonald, 1999) e.g. concluding μb > μa when it is less than.

Traditionally an α of 0.05 has been used (Sokal and Rohlf, 1995). Selecting a value for α is subjective and requires researcher judgment (Lauckner, 1989; Neyman and Pearson, 1928). Appropriate α values are determined in part, by the seriousness of committing different types of errors (Carmer, 1976). Shrader-Frechette and McCoy (1992) discuss preferences for committing certain types of errors in the context of different types of scientific rationality. They suggest that type I error prevention is preferred for pure science (preference for failing to acknowledge a truth over accepting a falsehood) and type II error prevention is preferred for applied science (preference for avoiding harm or loss of benefit) (Shrader-Frechette and McCoy, 1992).

Agriculture is an applied science and therefore may be expected to use larger α-values (risking type I error) in order to reduce type II errors. In the extreme Carmer (1976) argues that farmers use an α of 1.0 to assess variety trials in order to prevent type II errors. As such, farmers willingly commit type I errors because this does not affect production but will unknowingly commit type III errors. Setting α-values for on-farm trials will depend on the type of trial, the end use of the information, and the larger context of the trial. If trials are comparing treatments
that provide benefits at low costs to farmers then larger $\alpha$-values may be used (e.g. variety trials
for non-proprietary seed) but if costs are incurred (e.g. fertilizer trials) more conservative $\alpha$-
values should be used.

Greenwood (1993) argues that power calculations are as important as significance tests.

Power is the probability that a given statistical test will correctly reject the null hypothesis when
it is false and is defined as: $1 - (\beta + \gamma)$ (MacDonald, 1999). Power can be assessed for many
statistical tests (Zar, 1999) but for some testing procedures such as mixed models it is still an
area of ongoing research (Castelloe, 2000) and for others such as non-parametric tests simulation
techniques are used (Thomas and Juanes, 1996). Zar (1999) suggests that power may be
assessed after doing a study, in contrast Anderson et al. (2001) argue that it is only valid to assess
power prior to implementing studies.

Agricultural studies often analyze a number of variables from a single study. Performing
a series of tests to analyze more than one variable will result in $\alpha$ becoming inflated (Zar, 1999).
For a suite of 10 variables analyzed at an $\alpha$ of 0.05, a significant difference will be detected for at
least one variable 40% of the time even if no difference exists (Rice, 1988). Mathematically for
$n$ comparisons the probability of making at least one type I error is: $1 - (1-\alpha)^n$. A number of
adjusted rejective tests are available for such circumstances (Westafall et al., 1999), but it would
be inappropriate to apply such tests to all analyzes in a monograph (Rice, 1988) so researchers
must decide what constitutes independent analysis. Alternatively, a multivariate analysis of
variance (Wilks, 1932) may be used which has the additional advantage of considering the
correlations among variables (Zar, 1999).

Typically agricultural studies compare more than two treatment levels at a time (Cramer
and Walker, 1982). Comparison of multiple treatment levels for a single variable (multiple
comparisons) also requires test procedures that do not inflate $\alpha$-values (Zar, 1999). Day and Quinn (1989) review a number of these unplanned multiple test procedures and provide suggestions for appropriate procedures given data with different characteristics (unequal sample sizes, unequal variances, use of a control, etc.). The use of unplanned multiple comparisons in agricultural science has often been inappropriate and uncritical (Lauckner, 1989). Multiple comparison tests fail to interpret experimental data properly because they disregard the reasons behind choosing experimental treatments (Mead, 1986). Researchers should always attempt to outline questions they wish to test and conduct appropriate planned comparisons (Day and Quinn, 1989).

Increasing sample size increases power (Sokal and Rohlf, 1995). Alternatively certain statistical tests differ in their power. The power of non-parametric techniques are often less affected by violation of the assumptions of parametric tests (Sokal and Rohlf, 1995). Non-parametric tests tend to have superior power for a number of distribution types (MacDonald, 1999). Following from this a significant result achieved using parametric techniques for data that violate assumptions (normality, equality of variance) is not suspect because the penalty paid for using parametric techniques is reduced power not an inflated $\alpha$. Caution is advised because simulations indicate that in some cases $\alpha$ inflation may occur (MacDonald, 1999).

Farmers are not concerned with the mechanics of statistics, they desire results that are biologically meaningful to them. Given a large enough sample size statistical significance can almost always be demonstrated, and conversely a small-sample size may fail to detect biologically important differences (Thomas and Juanes, 1996). When biologically significant ranges are not known power analysis should be conducted for small, medium, and large differences and gives suggestions as to the magnitude of differences (small, medium, large) for
different types of data (Cohen, 1988). Generally there is a trend to move away from statistical
hypothesis testing (Anderson et al., 2000; Jones and Matloff, 1986; Nester, 1996; Yoccoz, 1991)
and towards the estimation of effects, their sizes, and confidence intervals (Burnham and
Anderson, 2002). This can be viewed as an attempt to move statistical inference away from the
subjective practice of hypothesis testing and towards evaluating biological significance.

As mentioned parametric analytical procedures are based on assumptions. For instance
analysis of variance procedures assume: 1) normality of errors, 2) equality of variance, 3)
independence, 4) non-additivity (Sokal and Rohlf, 1995). Analysis of covariance makes
additional assumptions: 1) covariate is linearly related to the dependent variable, 2) the covariate
is not related to the treatment effect, 3) the covariate is free of sampling error. Different test
procedures make different assumptions and researchers must review these when conducting
analyses. Cochran (1947) reviewed the consequences of violating assumptions for analysis of
variance statistical procedures (Table 1).

<table>
<thead>
<tr>
<th>Violation</th>
<th>Consequences</th>
<th>Difference between estimates</th>
<th>Accuracy of estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis F-tests</td>
<td>Loss of power</td>
<td>Slightly less efficient</td>
<td>Unbiased</td>
</tr>
<tr>
<td>Normality of errors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equality of variance of errors</td>
<td>Loss of power</td>
<td>Less efficient</td>
<td>Possible bias</td>
</tr>
<tr>
<td>Independent observations</td>
<td>Substantial impaired</td>
<td>Biased</td>
<td>Substantial bias</td>
</tr>
<tr>
<td>Additivity</td>
<td>Biased</td>
<td>Biased</td>
<td>Relatively unbiased</td>
</tr>
</tbody>
</table>

Table 1. Consequences of violating analysis of variance assumptions

Source: Adapted from Cochran (1947)
Zar (1999) and Sokal and Rohlf (1995) describe a number tests that can be applied to assess whether data depart from these assumptions and recommend ameliorative measures including a large number of data transformations. Bartlett (1947) presents less common transformations. Missing data is a problem in on-farm research and plant science experiments are commonly unbalanced (Spilke et al., 2005). Data may be missing because of discarded diseased plots (Brandjes and Lauckner, 1997), unanswered interview questions (Day and Evers, 2002; Malton, 1983), poor management of farm replicates (Fielding and Riley, 1998), and lost samples (Brandjes and Lauckner, 1997). Incomplete blocks are also common where not all farms receive the same treatments (Snapp, 2002) due to farm conditions and/or farmer preference.

The pattern of missing data determines what ameliorative steps should be taken. If large amounts of missing data are concentrated on a few variables or locations then data deletion is appropriate, whereas if missing data are few and randomly distributed data estimation measures may be used (McCune and Grace, 2002). For analyses based on ordinary least squares the sums of squares must be adjusted based on estimated values (Glen and Kramer, 1958). The use of maximum likelihood approaches (discussed below) is considered a more flexible and powerful approach to missing data that requires fewer assumptions (Enders, 2001).

Indices are a way to summarize data by combining several measures into a single number (Weigelt and Jolliffe, 2003) and are useful for interpreting intercropping, a common practice among resource-poor farmers (Chambers and Ghildyal, 1985). Indices used for intercropping include the land equivalent ratio (Mead, 1986; Oyejola and Mead, 1981; Riley, 1984), monetary equivalent ratio (Adetiloye and Adekunle, 1989), and numerous plant competition indices (Adetiloye and Adekunle, 1989; Weigelt and Jolliffe, 2003). Indices have been criticized as concealing real data (Mead, 1986; Weigelt and Jolliffe, 2003) but remain useful.
Analysis of variance has become a standard technique for analyzing agricultural experiments (Lauckner, 1989). It is a special form of regression in which factors and their levels are nominal in scale (e.g. blocks, varieties, etc.). Analysis of variance partitions total variation in the model (sums of squares) among the factors specified in the model. Hypothesis tests are then conducted for factors based on the ratio (F-stat) of the factor’s variance to the error variance, which is subsequently related to a probability on the basis of the F-distribution. When there is more than one factor being analyzed and these factors are crossed to each other the analysis of variance includes interactions. If interactions are non-significant their variance and degrees of freedom can be pooled with the error term (Bancroft, 1964) because under such circumstances the interaction term and error term are estimates of the same population variance (Zar, 1999). Sokal and Rohlf (1995) provide a dichotomous key based on Bancroft (1964) for deciding when to pool.

Analysis of variance can be modified by measuring covariates and including them in the model. When covariates are included, the analysis of variance is then termed an analysis of covariance, and such analysis is common in agriculture (Stroup, 1989). Analysis of covariance is a combination of linear regression and analysis of variance in which the variable not included in the experimental design is used for regression. If the covariate explains a significant amount of variation then values for hypothesis tests and parameter estimates will be adjusted accordingly (Wildt and Ahtola, 1978).

In on-farm research there is debate about how to treat interactions between farms and treatments. Gomez and Gomez (1984) view explaining interactions as the most important aspect whereas, Stroup et al. (1993) argue that the interaction is not of interest because it merely indicates that treatment differences vary randomly by farm as expected. These divergent
opinions are in some ways based on the contrast between the notions that on-farm trials should develop local technologies (interaction important) versus general technologies (interaction unimportant). Interactions between two factors can be assessed graphically (Nelson, 1988) and are either cross-over interactions (differences in rank order) or amplitude interactions (deviations from additivity) (Truberg and Huhn, 2000). If there are no within farm replicates it is not possible to conduct a hypothesis test for the farm-by-treatment interaction because it is equivalent to the error term (Littell et al., 1996). Tukey’s test for non-additivity may be applied to detect interactions in such cases (Tukey, 1949). Alternatively, Hildebrand (1984) proposed modified stability analysis, based on Eberhart and Russell’s (1966) work, to assess crop variety by farm interactions.

To explain on-farm experiment interactions Cady (1991) proposes regression against environmental variables with the order of variable inclusion based on researcher knowledge and the number of variables kept small relative to the number of farm replicates. Mutsaers (1991) suggested that order of inclusion should be: 1) physical variables not controlled by the farmer (e.g. soil), 2) farmer controlled physical variables (e.g. weediness), 3) non-physical variables (e.g. gender), and 4) composite variables (multiple unknown causes).

In the case of subsamples one factor is contained within another factor rather than crossed to it. This results in a nested design (Hurlbert, 1984) for which hypothesis test error terms must be appropriately defined. For instance if you have farms at each of two locations and farms are not replicated at (crossed with) each location then farm is nested within location. Testing the location effect requires that the farm within location mean square be used as the F-stat denominator and not the model residual/error term.
Non-parametric statistical methods do not make the same assumptions about population distributions and variances as parametric tests. Given that the variability of on-farm trials is more complex (Stroup et al., 1993) data are less likely meet normality and variance assumptions. Also, Stevens (1946) has argued that only non-parametric statistics can be used to analyze nominal and ordinal data (ranking, scoring, etc.). In contrast Zar (1999) finds that ordinal data can be analyzed using parametric techniques although ordinal data is more likely to violate normality assumptions. Many statistical texts (e.g. Sokal and Rohlf, 1995; Zar, 1999) provide formulas for calculating non-parametric statistics (see Table 3).

Table 3. Non-parametric tests and their parametric equivalents

<table>
<thead>
<tr>
<th>Non-parametric tests</th>
<th>Equivalent parametric test</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilcoxon Rank Sums Test</td>
<td>T-test</td>
<td>(Wilcoxon, 1945)</td>
</tr>
<tr>
<td>Wilcoxon Signed Rank Test</td>
<td>Paired T-test</td>
<td>(Wilcoxon, 1945)</td>
</tr>
<tr>
<td>Kruskal-Wallis ANOVA</td>
<td>One-way ANOVA</td>
<td>(Kruskal and Wallis, 1952)</td>
</tr>
<tr>
<td>Friedman ANOVA</td>
<td>Randomized block design</td>
<td>(Friedman, 1937)</td>
</tr>
<tr>
<td>Spearman Rank Correlation</td>
<td>Pearson Correlation</td>
<td>(Spearman, 1904)</td>
</tr>
</tbody>
</table>

These tests use ranks instead of measured values, which makes their parameter estimates difficult to interpret and as such these techniques are generally used for hypothesis testing only.

Mixed models use a combination of random and fixed factors (Zar, 1999). A factor in a model is a fixed factor if its levels have been intentionally selected for experimentation whereas a factor is random if the levels of the factor are a random sample from a population (Littell et al., 1996). Interaction factors are considered random if one or more the interaction factors is random. Mixed models are particularly useful for analyzing on-farm trials (Littell et al., 1996) because the farm factor refers to a population of farms. In an on-farm trial if farm is considered a fixed effect the inference space of the analysis is limited to farms included in the study instead
of a population of farms (Littell et al., 1996). In mixed model analysis of variance (type III ANOVA) using ordinary least squares, assessment of significance can be performed by modifying F-statistic computation with fixed factor variance divided by the interaction effect variance (Zar, 1999).

Alternatively a maximum likelihood approach may be used. Currently, there is a trend towards the use of maximum likelihood techniques over traditional least squares techniques (Burnham and Anderson, 2002). Maximum likelihood approaches aim to find values for model parameters (e.g. variances) that make the observed data most likely/probable. This requires: 1) development of maximum likelihood equations/estimators to find model parameter values; 2) use of computational algorithms must be used to search dataset parameter space to find solutions to these estimators; and 3) testing hypotheses regarding the maximum likelihood estimates.

There are many types of maximum likelihood estimators (e.g. maximum likelihood, restricted/residual maximum likelihood, quasi-maximum likelihood, etc.). These estimators can be defined in terms of different types of distributions (Myers and Cadigan, 2001) (e.g. binomial, normal, gamma, etc.). Hartley and Rao (1967) provided a maximum likelihood procedure (ML) for producing maximum likelihood estimates in mixed model analysis. Work by Paterson and Thompson (1971) and Corbeil and Searle (1976) developed the restricted maximum likelihood procedure (REML) for mixed models which unlike the ML procedure accounts for loss of degrees of freedom due to fixed effect estimation (Harville, 1977). There are many iterative numerical algorithms that can be used to compute maximum likelihood estimates with some being estimator specific while others can be generally applied (Harville, 1977). Some common algorithms are: EM (Expectation and Maximization) algorithm (Dempster et al., 1977) and the subsequently derived ECM (Expectation and Constrained Maximization) (Meng and Rubin,
1993), Newton-Raphson algorithms (Lindstrom and Bates, 1988). These algorithms efficiently
search parameter space for a solution, using rate of change techniques (derivation).

In mixed models the squared standard error of a contrast cannot be expressed as a linear
model of a single variance component. As such approximations for degrees of freedom are used
(Spilke et al., 2005). Satterwaite (1946) developed an degrees of freedom approximation for
balanced models employing least squares estimation. This technique was then further
generalized so that it could be applied to unbalanced models (Fai and Cornelius, 1996) and later
improved by Kenward and Roger (1997) to correct for bias.

Models are composed of a number of parameters with each parameter explaining part of
the variation in the data. In on-farm trials these may be factors such as: farm, year, location,
treatment, environmental covariates, etc. Selection of appropriately paramitized models for is is
important and many techniques are available (Burnham and Anderson, 2002). Akaike’s
information criterion has become a popular statistic for model selection (Akaike, 1974).
Bayesian information criterion (BIC) are also available for model selection (Burnham and
Anderson, 2002). AIC is preferred for situations in which models consist of a number of small
tapering effects while BIC is preferred for models consisting of a few major effects (Burnham
and Anderson, 2004).

Statistical techniques for assessing spatial variation are largely from the field of
geosatistics and Vieira et al. (1983) review their application to agricultural research. Dale et al.
(2002) review a large number of spatial statistics and how they are mathematically related.
Multivariate techniques such as partial canonical analysis may be used to assess causation
between factors (e.g. yield and soil fertility) (Legendre and Legendre, 1998).
Conclusions

There are a wide variety of design and statistical analysis techniques that can be drawn on when conducting survey and/or trial work in on-farm research. Unlike laboratory or on-station research on-farm research must confront the practical constraints of working with farmers. These considerations strongly influence design. Statistical techniques for analyzing data are subsequently influenced by design. Generally, on-farm research uses more complex designs that require more flexible and generalized statistical methods. Design takes precedent over analysis and the numerous existing statistical analysis tools should be used as design necessitates. Mixed model analysis has been presented as one of the more flexible methods available. No review of experimental design and statistical analysis for on-farm research can be exhaustive. This review serves as a starting point for further reading.

9.0 Literature cited


CIMMYT. 1982. A design decision in on farm experiments. Farming systems newsletter 11:4-8.


experimental procedures Caribbean Agricultural Research and Development Institute
(CARDI), Trinidad.
Harville, D.A. 1977. Maximum likelihood approaches to variance component estimation and to
Hasel, A.A. 1938. Sampling error in timber surveys. Journal of Agricultural Research 57:713-
736.
systems research symposium. University of Arkansas and Winrock International Institute
for Agricultural Development, Fayetteville.
Journal 76:271-274.
Hildebrand, P.F., and F. Poey. 1985. On-farm agronomic trials in farming systems research and
extension Lynne Rienner, Colorado.
Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments.
Ecological Monographs 54:187-211.
Huxley, P.A. 1987. Agroforestry experimentation: separating the wood from the trees?
196:19-25.
Justesen, H.S. 1932. Influence of size and shape of plots on the precision of field experiments
Kalamkar, R.J. 1932. Experimental error and the field plot technique with potatoes. Journal of
Lafond, G.P., and F.C. Stevenson. <2004>. Improving on the efficiency of multi-factor and
multi-location agronomic experiments. Agriculture and Agri-Food Canada, Indian Head, SK.
Langton, S. 1990. Avoiding edge effects in agroforestry experiments; the use of neighbour-
Lauckner, F.B. 1989. Survey of the use of statistics in agricultural research journals. Tropical
Agriculture (Trinidad) 66:2-7.
Lawrence, P.R., B. Gerard, C. Moreau, F. Lheriteau, and A. Buertkert. 2000. Design and testing
of a GPS-based reflectometer for precision mapping of pearl millet total dry matter in the


Stroud, A. 1993. Conducting on-farm experiments Centro Internacional de Agricultural Tropical, Cali.


