

# Analyzing Ranking and Rating Data from Participatory On-Farm Trials

RICHARD COE

## Abstract

*Responses in participatory on-farm trials are often measured as ratings (scores on an ordered but arbitrary scale) or rankings (respondents are simply asked to order treatments). Usual analysis of variance and linear model based analyses are not appropriate for these data. Alternative analyses based on generalized linear models are described in this paper. These methods can be successfully used when trial designs are irregular—a common characteristic of participatory trials—and when covariates are measured on each plot or farm in order to identify genotype x environment interaction.*

## Introduction

Participatory methods have been widely adopted by researchers working on applied agricultural problems including crop breeding. This change in paradigm has implications for the design and analysis of methods used. Some of these are summarized in a companion paper by this author (see Coe 1, this proceedings). An assumption of this paper is that formal analysis of quantitative data collected systematically from trials is still an important part of the process. Without this it is difficult to see how the research activity can generate information relevant to anyone other than the small number of farmers directly involved. Breeders that have adopted participatory methods have generally recognized this, but have faced three difficulties:

1. The experimental designs used are often irregular in layout due to farmer participation (e.g., in choosing which varieties to test on their farms) or constraints arising from trials being located in farmers fields.
2. The focus of analysis often shifts from overall varietal selection to understanding the variation in varietal response across farms. This is genotype x environment interaction (GEI), where the E may include social or economic variables in addition to biophysical environments.
3. Much of the quantitative data collected may be ratings and rankings, for which the more usual methods of analysis may not be appropriate.

In this paper, I describe the analysis methods that simultaneously deal with each of these difficulties. I start by summarizing the nature of ranking and

rating data and reviewing approaches to analysis. Examples to illustrate the analyses are presented in the third section. In the fourth and fifth sections, I present a detailed discussion of an approach to analysis of ranking and rating data, respectively. The discussion in the final section highlights some outstanding problems and implications of the methods presented.

## Data and Analysis Types

The nature of the response variable is one factor which determines the type of analyses that can be conducted, whether formal or informal. It is therefore important to understand exactly how the data were collected and what the numbers represent.

### Continuous

Quantities such as crop yield can be measured on a continuous scale, e.g., in kg/m<sup>2</sup>. The numbers have the property that “2 really is the average of 1 and 3”, making many common statistical procedures appropriate. Such quantities may be on a “ratio” or “interval” scale, the difference depending on whether the scale has a real zero. For example, a yield of 1 t/ha is 50% of a 2 t/ha yield, but a temperature of 10°C is not 50% of 20°C, as the zero for temperature is arbitrary.

### Scores or rating

Here I refer to data that is recorded on a scale from poor to excellent, or less than enough to more than enough. The categories used are often given numerical labels, such as 1, 2, 3, 4, 5. These are called scores or ratings and such a scale is also described as ordered categorical. The labels are arbitrary. An

observation of 3 is higher than an observation of 2, but we cannot say that it is better by the same amount that an observation of 5 is better than one of 4. An analysis would ideally order the data without using the actual numerical label to ensure that the results are the same regardless of which labels are used.

### Binary

Data recorded with just two categories are common, e.g., yes or no, dead or alive, acceptable or not acceptable. Analysis is based on the frequency with which the categories occur. Analysis methods for these data are widely described in the literature but are not discussed further here.

### Ranks

In many investigations of preference, data are collected by asking respondents to rank alternatives. The options available are placed in order without any attempt to describe how much one differs from another or whether any of the alternatives are, for example, good or acceptable. We might have variety A ranked above B, which is ranked above C, yet none of the three are considered good. The data would look the same in the case where a respondent placed them in the same order, but one, two, or all three were acceptable.

Other scales may be hybrids of these.

The steps in analyzing any data set can be summarized as:

1. Define the analysis objectives. These drive the rest of the analysis. It is impossible to conduct a good analysis of data without clear objectives. Often the key graphs and tables can be defined at this stage, even without the results with which to fill them in.

2. Prepare the data. Data sets will have to be entered and checked, suitable transformations made (e.g., to dry weight per unit area), relevant information from different sources (e.g., farm household data and plot level yields) extracted to the same file, and so on.
3. Exploratory and descriptive analysis. The aim is to summarize the main patterns and notice further patterns that may be relevant. This step is only covered briefly in this paper, as the methods used will depend on the context in which the analysis is carried out and the audience at which the results are aimed.
4. Formal statistical analysis. The aim is to add measures of precision and provide estimates from complex situations.
5. Interpretation and presentation.

Iteration between the steps will be necessary. Training materials by Coe et al (2001) provide more information on analysis of experiments. Some comments on the roles of these steps in analysis of participatory experiments are given in the companion paper (Coe 1, this proceedings).

A common objective in the analysis of many participatory breeding trials is to understand the nature of variation in responses from different farmers. Many researchers report that participatory on-farm trials give highly variable results, making interpretation difficult. Certainly if a standard analysis aimed at identifying differences in variety means is conducted, the result may well be a very high residual variation with correspondingly large standard error or variety differences, implying only vague

knowledge about the relative performance of the entries. However, the variation can often be understood as genotype x environment interaction (GEI). The environment in which a participatory trial takes place is heterogeneous. Though a researcher may have full control over a trial, there will be many sources of hidden variation, including social or economic factors, as well as the more usual biophysical definitions of environment. For example, male and female farmers may assess varieties differently, or ratings may depend on the level of market integration of a farmer. The analyses must therefore be able to identify and describe these GEIs. When this occurs, the results are often the most useful output of the trial because they allow recommendations to be adjusted to particular local conditions

A spreadsheet package such as Excel is good for much of the descriptive analysis. Its flexible facilities for data selection and transformation, tabulation, and graphics are useful. However, dedicated statistical software is needed for the analyses described here—they cannot be performed in Excel. There are several packages with approximately equivalent facilities. All examples given in this paper use Genstat (2000)—I find it often the easiest to understand, particularly since methods for different problems can be addressed using a similar sets of commands. The key commands used to produce each analysis are included in the text with their output. SPSS is widely used by social scientists but is not particularly useful for the analyses described here.

## Examples to Illustrate Analysis Methods

### 1. Soil fertility under agroforestry in Malawi

This is not a breeding trial but is included because the design is typical of many participatory on-farm trials. Three soil fertility strategies are compared over a number of years:

- g Mixed intercropping of maize and gliricidia
- s Relay planting of maize and sesbania
- c The control of continuous maize

Forty-one farmers each compared the control with one or both of the other treatments. Crop yield is the response of interest. A number of covariates were measured at the plot or farm level to help understand the reasons for variation across farms.

### 2. Maize varieties in Zimbabwe

This was a “baby” trial.<sup>1</sup> Twelve maize varieties were compared. A total of 146 farmers in 25 different sites took part, each farmer testing 4 of the 12 varieties. The varieties tested by each farmer were chosen by the researcher. Some household and field covariates were recorded. The actual crop yields obtained were not available for analysis, so the examples described here use simulated yield data but the original field design.

### 3. Maize varieties in Kenya

In this baby trial, 18 varieties of maize were compared, 2 of which were local controls. Twenty-nine farmers

participated, each planting 2 replicates of all 18 entries. Crop performance was rated on a scale of 1, 2, 3, 4, 5. Gender of the respondent and farm size were also recorded.

## Analyzing Ratings or Scores

### Example 1

The crop yields in Example 1 were measured in tons per hectare; however, I have converted them to scores to illustrate the method of score analysis. The conversion is “exact” (i.e., the scores are those that farmers would give if asked to assess yield and are without error) so that for illustration purposes the results can be compared with those obtainable from actual yields. Scores were allocated as follows:

Yield	Score	Label
$y < 1$	1	Poor
$1 \leq y < 2$	2	OK
$2 \leq y < 3$	3	Good
$3 \leq y$	4	Excellent

We could just tabulate frequencies as:

```
TABULATE [PRINT=nobs;
CLASSIFICATION=trt,score98; MARGINS=no]
score98
```

score98	Nobsrvd			
	poor	ok	good	excellent
trt				
c	9	13	6	3
g	5	15	10	9
s	3	7	9	5

<sup>1</sup> The mother-baby trial design comprises a central researcher-managed “mother” trial, which tests all varieties, and farmer managed “baby” trials, which test a subset of the varieties from the mother trial.

This is informative. For example, the mode of the distribution for treatment g is deemed "OK". This shifts to "good" for treatment s. For treatment c the mode is also OK but the frequencies of other scores suggest that g is better than c.

This type of analysis has obvious drawbacks:

1. It is difficult to know how to handle more complex patterns, e.g., two modes.
2. It seems to ignore some of the structure in the data. For example, we have not used the fact that each farmer rates two or three treatments.
3. It is not obvious how the analysis could be extended to deal with more complex problems such as identifying and describing the effects of covariates to describe GEL.
4. It is not obvious how to formalize the analysis so that we can give measures of uncertainty, i.e., standard errors, confidence intervals, or statistical hypothesis tests.

A common approach is to treat the scores as quantities measured on a continuous scale. Means can then be calculated (see below) and all methods of analysis of variance, regression, and related modeling can be tried.

	Mean	Variance
trt		
c	2.097	0.8903
g	2.590	0.9852
s	2.667	0.9275

There two reasons to feel uncomfortable about this. The first is that many of the assumptions of analysis of variance or linear regression modeling may well be inappropriate, given the limited range of the observations. A critical assumption is that the variance between observations

of the same treatment is constant across treatments. This is commonly not the case, with the extreme entries showing less variation in score than those with a mean of 1 or 2.

The second reason is that the method makes some assumptions about the scores that may not be appropriate. For example, is the average of "poor" and "good" really "OK"? The seriousness of this objection is obvious when it is realized that the scores 1, 2, 3, and 4 are just labels, but the results critically depend on which label is given. If we used 0, 1, 5, 100, for example, the results would look very different, yet logically speaking these labels are equally acceptable.

There are situations in which both these objections are unimportant and so a useful analysis can be made. However, we would prefer an approach that is theoretically more sound and more robust, and applicable in a wider range of cases.

A second approach is to dichotomize the response, i.e., change it from a 4-level to a 2-level scale. For example, we could group poor and OK together and good and excellent together to give a measure with just two possible values. There are well established methods for analyzing such data, including models (e.g., logistic regression) that allow the effects of complex arrangements of covariates to be disentangled, and even methods (generalized linear mixed models) that allow random effects to be incorporated, such as the REML analysis of continuous data (Coe 1, this proceedings). However, this approach is also unsatisfactory. If the variable is originally measured on a 4-point scale

and we reduce it to a 2-point scale, then we must be losing information.

Valid methods have been developed which use the all information without making unreasonable assumptions and can model the effect of covariates. In order to understand the model, we look first at the data for just two treatments, g and c, and forget that the observations are paired by farmer. The data are thus the following frequencies:

Treatment	Poor	OK	Good	Excellent
c	9	13	6	3
g	5	15	10	9

If we combine the top three categories, the data is reduced to a 2 x 2 table:

Treatment	Poor	OK + Good + Excellent
c	9	22
g	5	34

From these tables it looks as though g is better than c. A higher proportion of the plots are in the OK + good + excellent category. A common measure of this association is the odds ratio, O, or log odds ratio,  $\log(O) = L$ , which can be calculated as follows:

$$O = \frac{\text{odds on g high}}{\text{odds on c high}} = \frac{34/5}{22/9} = 2.78$$

$$L = \log(2.78) = 1.02$$

If we “cut” the categories differently, combining poor and OK, the following data are produced:

Treatment	Poor + OK	Good + Excellent
c	22	9
g	20	19

From this table  $O = 2.32$ , and  $L = 0.84$ .

A third cut is possible, combining poor, OK, and good to give:

Treatment	Poor + OK + Good	Excellent
c	28	3
g	30	9

From this table,  $O = 2.80$  and  $L = 1.02$ .

In this case the O values are similar for each cut. If we make the assumption of such proportional odds with a constant value of O, then its value and standard error can be estimated without choosing any particular cut. In Genstat the calculations are made using the regression modeling commands. Note that the data have to be arranged so that there is a response variable for each possible response category. The variable for each score contains the number of plots that had that score.

```
print treat, s1, s2, s3, s4

      treat   s1    s2    s3    s4
      c      9.000 13.00  6.000  3.000
      g      5.000 15.00 10.000  9.000

model [dist=multinomial; yrel=cumulative; link=logit]
s1, s2, s3, s4 fit [p=e, a] treat
```

```
***** Regression Analysis *****
*** Estimates of parameters ***

                                antilog of
                                estimate
estimate s.e. t(*) estimate
Cut-point 0/1 -0.927 0.367 -2.53 0.3956
Cut-point 1/2 0.948 0.367 2.58 2.581
Cut-point 2/3 2.161 0.438 4.93 8.680
treat g      0.932 0.452 2.06 2.539
```

\* MESSAGE: s.e.s are based on dispersion parameter with value 1

Parameters for factors are differences compared with the reference level:

Factor	Reference level
treat	c

\*\*\* Accumulated analysis of deviance \*\*\*

Change	d.f.	deviance	mean deviance	deviance ratio
+ treat	1	4.37545	4.37545	4.38
Residual	2	0.16035	0.08018	
Total	3	4.53580	1.51193	

\* MESSAGE: ratios are based on dispersion parameter with value 1

The analysis of deviance is interpreted similarly to an analysis of variance, comparing the deviance with a chi-squared distribution to judge the importance of the effect. In this case there seems to be a significant treatment difference.

The parameter estimate `treat g` measures the difference between treatments `g` and `c`. The estimate 0.932 is the log odds ratio =  $\log(\text{odds of } g \text{ being high vs. low} / \text{odds of } c \text{ being high vs. low})$ . Here high and low refer to being above and below a cut point, respectively, in the ordered set of scores (it doesn't matter which cut point, as the model constrains this odds ratio to be the same for any choice of cut point).

The value of 0.932 for the log odds ratio means that the odds ratio is  $\exp(0.932) = 2.539$ . This is similar to the average of the

three odds ratios found directly from the data. The standard error can be used to test the hypothesis of no difference between `g` and `c` (log odds ratio of 0) or to give a confidence interval for the log odds ratio or odds ratio. The cut-point parameters listed by Genstat do not have a useful interpretation in this case.

Now we analyze the whole data set using the same ideas and include a term for `farm` to account for the fact that each farmer is evaluating two or three plots. There is a row of data for each plot and a column for each possible score (poor, ok, good, excellent, or 1, 2, 3, 4), which are now assigned the names `s98[1]`, `s98[2]`, `s98[3]`, and `s98[4]`. The data value is again the number of plots that were given that score, but now this value is simply 0 or 1, with a single 1 given in each row. A small part of the data is shown:

```
print name, trt, s98[1..4], score98; 10; decimals=0
```

name	trt	s98[1]	s98[2]	s98[3]	s98[4]	score98
Chakame	g	0	0	0	1	excellent
Chakame	s	0	0	1	0	good
Chakame	c	1	0	0	0	poor
Thobola	g	0	1	0	0	ok
Thobola	s	1	0	0	0	poor
Thobola	c	0	1	0	0	ok
Adisani	g	0	1	0	0	ok
Adisani	c	1	0	0	0	poor
Majoni	g	0	0	0	1	excellent
Majoni	s	0	0	0	1	excellent
.	.	.	.	.	.	.

```
model
[dist=multinomial; yrel=cumulative; link=logit]
s98[1..4]
fit [p=*] name
add [p=a] trt
```

\* MESSAGE: Term name cannot be fully included in the model because 2 parameters are aliased with terms already in the model

```
(name Komwa(died 97)) = 0
```

```
(name Lipenga(died 98)) = 0
```

```

***** Regression Analysis *****

*** Estimates of parameters ***

                                antilog of
                                estimate
estimate s.e. t(*) estimate
Cut-point 0/1  1.90  2.00  0.95  6.716
Cut-point 1/2  7.23  2.27  3.18  1382.
Cut-point 2/3 10.72  2.43  4.41 45305.
name Belo     4.04  2.56  1.58  56.75
name Bisiwiki 0.00  2.76  0.00  1.000
name Chakame  5.63  2.51  2.24  278.9
name Chimimba 0.97  3.57  0.27  2.638
.
.
name White    5.56  2.51  2.21  259.7
trt g         3.598  0.770  4.67  36.51
trt          2.722  0.786  3.47  15.21

```

\* MESSAGE: s.e.s are based on dispersion parameter with value 1

Parameters for factors are differences compared with the reference level:

Factor name	Reference level
trt	c

\*\*\* Accumulated analysis of deviance \*\*\*

Change	d.f.	deviance	mean deviance	deviance ratio
+ name	38	115.468	3.039	3.04
+ trt	2	30.218	15.109	15.11
Residual	51	105.977	2.078	
Total	91	251.663	2.766	

\* MESSAGE: ratios are based on dispersion parameter with value 1

The analysis of deviance is interpreted in the usual way using a chi-squared distribution to assess the size of contributions. A deviance of 30.2 with 2 df confirms that treatment is having an obvious effect on the ratings.

The parameter estimates for each farmer are uninteresting since they only reflect the fact that farmers can differ in the mean rating given, whereas the estimates

for the treatments are important because they give the quantitative summary of the ratings. In this example the control treatment c is the baseline from which the others are measured. The important results are shown in the table below. For comparison, analysis of the actual yields using a similar method (linear model fitting farmer + treatment effects) is also shown. (For details see Coe 1, this proceedings.) Remember the scales are different. We cannot hope to recover information on actual yield per hectare from data that has simply been recorded as poor, OK, etc. What are important are the differences and similarities between treatments, which are revealed by this analysis.

Treatment	Rating		Yields		Scaled yields <sup>†</sup>	
	Log odds ratio	se <sup>‡</sup>	Adjusted mean	se	Adjusted mean	se
g	3.60	0.77	2.62	0.15	3.60	0.53
s	2.72	0.79	2.37	0.17	2.68	0.61
c	0.00	-	1.64	-	0.00	-

<sup>†</sup> Yield means scaled to match the log odds ratio scale.

<sup>‡</sup> se is the standard error of the difference from c.

When the scales are aligned, the results of the analyses are remarkably similar. The se values for the rating data are higher because ratings contain less information than actual yields.

The value of the analysis becomes clear when we start to look at differences between groups of farmers or try to understand the effect of covariates. For example, slope2 is a factor which classifies farms into flat or sloping categories. The variate cec is related to soil fertility. It was hypothesized that g would perform relatively better on flat land and that both g and s would be

superior to c when cec is low. These are investigated in the following table:

\*\*\* Accumulated analysis of deviance \*\*\*

Change	d.f.	deviance	mean deviance	deviance ratio
+ cec	1	1.945	1.945	1.94
+ slope2	1	8.959	8.959	8.96
+ trt	2	6.259	3.129	3.13
+ name	37	133.823	3.617	3.62
+ cec.trt	2	2.087	1.043	1.04
+ slope2.trt	2	2.543	1.271	1.27
Residual	44	90.606	2.059	

There is no clear evidence for either slope or cec showing an interaction with treatment.

**Example 3**

In this example, the performance of 18 varieties was rated on a scale of 1 (poor) to 5 (excellent). Criteria were yield, cob size, cob filling, and overall assessment. The design was straightforward: 29 farmers each evaluated the 18 varieties from 2 plots of each. Simple descriptive statistics can therefore give a useful summary of some of the characteristics. For example, the graph below only shows the frequency of responses for the overall

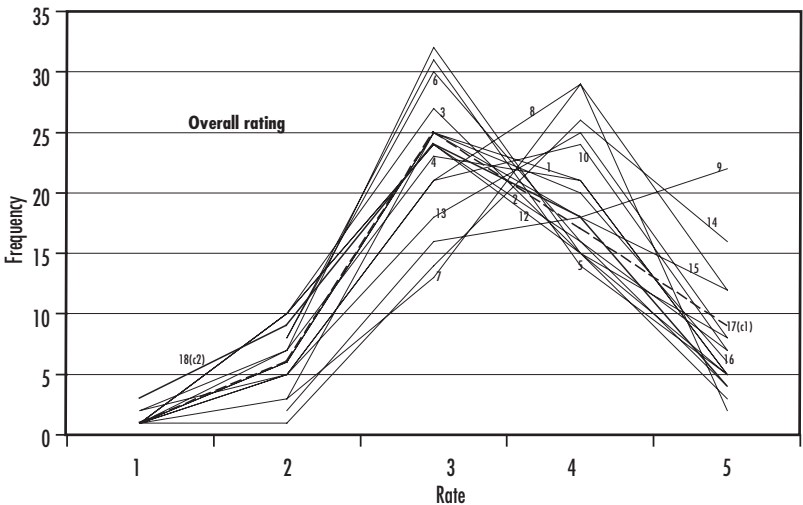
rating of each of the 18 varieties. Varieties 17 and 18 have been highlighted because they are local checks.

The varieties seem to fall into 2 main groups—those with a mode at 3 and those with a mode at 4—with entry 9 standing out by being rated higher than all the others.

There a number of reasons why the modeling analysis is still worthwhile:

1. It provides simple concise summaries with measures of precision.
2. It makes the inclusion of covariates straightforward. In this case both farm size and gender of the respondent have been recorded.
3. It simplifies the comparison of the ratings under different criteria.

The analysis follows a similar pattern to the previous example. Note that the layout with two replicates per farm can be explicitly included in the analysis if sensible. Here I have assumed the two replicates correspond to two blocks on each farm. Farms are distinguished by the factor IDNO and blocks within farmers by REP.



```

model
[dist=multinomial;yrel=cumulative;link=logit]
overall[]
fit [p=*]
add [p=*] IDNO
add [p=*] IDNO.REP
add [p=a,e] ENTRY
    
```

Parameters for factors are differences compared with the reference level:

Factor	Reference level
IDNO	1
ENTRY	c1

\*\*\* Accumulated analysis of deviance \*\*\*

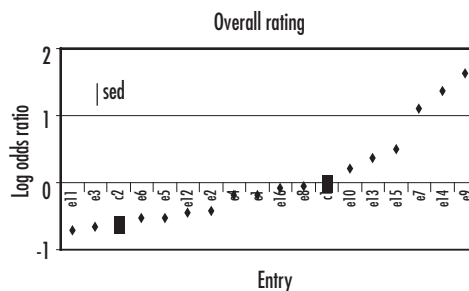
Change	d.f.	deviance	mean deviance	deviance ratio
+ IDNO	28	342.309	12.225	12.23
+ IDNO.REP	29	82.354	2.840	0.21048
+ ENTRY	17	123.623	7.272	7.27
Residual	966	2189.420	2.266	
Total	1040	2737.706	2.632	

\*\*\* Estimates of parameters \*\*\*

	estimate	s.e	t(*)	antilog of estimate
Cut-point 0/1	-7.333	0.609	-12.05	0.0006534
Cut-point 1/2	-4.558	0.541	-8.43	0.01048
Cut-point 2/3	-1.560	0.522	-2.99	0.2102
Cut-point 3/4	0.848	0.520	1.63	2.335
IDNO 2	0.170	0.645	0.26	1.185
.				
IDNO 29	-1.993	0.649	-3.07	0.1363
IDNO 1 .REP 2	-1.365	0.641	-2.13	0.2554
.				
IDNO 29 .REP 2	-0.319	0.652	-0.49	0.7271
ENTRY c2	-0.610	0.361	-1.69	0.5433
ENTRY e1	-0.182	0.359	-0.51	0.8338
ENTRY e2	-0.419	0.360	-1.16	0.6575
ENTRY e3	-0.653	0.361	-1.81	0.5206
ENTRY e4	-0.196	0.359	-0.55	0.8219
ENTRY e5	-0.530	0.361	-1.47	0.5883
ENTRY e6	-0.539	0.361	-1.49	0.5834
ENTRY e7	1.109	0.360	3.08	3.030
ENTRY e8	-0.049	0.359	-0.14	0.9523
ENTRY e9	1.625	0.365	4.45	5.078
ENTRY e10	0.223	0.358	0.62	1.250
ENTRY e11	-0.701	0.361	-1.94	0.4963
ENTRY e12	-0.438	0.360	-1.22	0.6453
ENTRY e13	0.377	0.358	1.05	1.458
ENTRY e14	1.380	0.362	3.81	3.974
ENTRY e15	0.510	0.358	1.43	1.666
ENTRY e16	-0.078	0.359	-0.22	0.9248

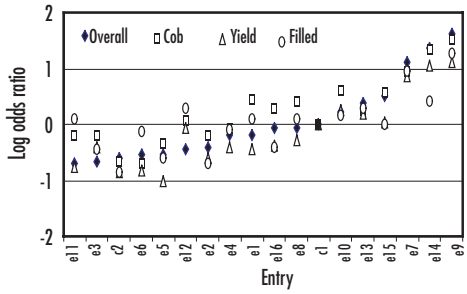
\* MESSAGE: s.e.s are based on dispersion parameter with value 1

The analysis of deviance suggests that there are large differences between the entries; the parameter estimates summarize these. Remember the estimates are log odds ratios that describe the chance of being in a high response category rather than a low one for each entry compared with the baseline. I set up the data files so that the baseline is the first local check, c1. A simple graph reveals the patterns:



Apart from entries 9, 14, and 7, there is a continuous spread of ratings of the varieties rather than any clear groupings, with one of the local checks occurring towards the lower end of the spread and the other towards the upper end.

The ratings on each criterion can now be compared by repeating the analysis and putting the log odds ratio for each on the same graph. The pattern is much the same for each criterion. Of the three best performing entries, 14 does least well for cob filling.



There are two covariates of interest recorded: the gender of the respondent and farm size. The question of interest is whether males and females rate the entries differently, and whether the relative ratings depend on farm size.

```

model
[dist=multinomial;yrel=cumulative;link=logit]
overall[]
fit [p=*

```

MESSAGE: ratios are based on dispersion parameter with value 1

Neither of the covariates show any interaction with entry; thus, we can conclude that the overall rating of varieties is much the same for males and females and does not have any linear relationship with farm size. The effect of farm size could perhaps be investigated further, for example, by putting farms

into a few (three or four) size categories. This approach removes the assumption of a linear relationship between farm size and the log odds ratios.

### Analyzing Ranks

At first glance, data from ranks look similar to rating data. Like ratings, the observations are integers from a limited range, and we want to find out the same sort of information, i.e., are there consistencies in the rankings given to different treatments that can allow us to reach conclusions about which treatments consistently ranked high? However, there are some important differences from rating data that will emerge.

#### Example 1

To illustrate the method I have once again converted yield data from Example 1 to ranks. Each farmer compared two or three treatments. Ranks have been allocated exactly, so that the treatment with the lowest yield on each farm is given rank 1, the next lowest rank 2, and the third (if there is one) rank 3. There were no ties. A small part of the data is shown.

name	yield98	rank	trt
Adisani	1.449	2.000	g
Adisani	0.801	1.000	c
Belo	*	*	c
Belo	2.071	2.000	s
Belo	1.246	1.000	g
Bisiwiki	0.643	1.000	c
Bisiwiki	1.514	2.000	g
Chakame	0.761	1.000	c
Chakame	3.380	3.000	g
Chakame	2.142	2.000	s
Chimimba	1.943	1.000	g
Chimimba	*	*	s
Chimimba	*	*	c
Chinzeka	2.356	3.000	g
Chinzeka	1.477	1.000	s
Chinzeka	1.713	2.000	c

Simple displays of the data can be designed. For example, we can tabulate the number of farmers who rank each treatment as 1, 2, or 3.

```
TABULATE [PRINT=counts;
CLASSIFICATION=trt,rank; MARGINS=no] yield98
```

Rank	Count		
	1.000	2.000	3.000
trt			
c	24	6	1
g	9	16	14
s	6	12	6

Unknown count 14

Treatment g is ranked 3 more often than s, which is an indication that it is superior. But a difficulty is immediately clear: g is also ranked 2 and 1 more often than s. The problem arises from the fact that each farmer is only ranking the treatments he/she tests, and these are not the same for each. In the table above, when g = 2 we cannot tell whether g was the best out of 2 treatments or second best out of 3. Changing the ranking method to 1 = best does not help. Some authors suggest converting ranks to scores, but of course the problem cannot be fixed by any conversion that simply changes the ranks 1, 2, 3 to another set of numbers.

A more realistic summary comes from studying each treatment pair. If we take, for example, g and c, we can look at all of the farmers that compared these two, and calculate the proportion that ranked g higher than c.

Pair	No. of comparisons	No. with first of pair ranked higher than second	Proportion with first of pair ranked higher than second
g-c	31	28	0.903
s-c	21	16	0.762
g-s	24	16	0.667

This summary now correctly only relies on the rankings within each farm and is explicit about what is compared with what. Its shortcomings and the reasons for wanting a formal analysis are much the same as for the rating data. We need to put measures of precision on results and would like to extend the analysis to look at the effect of covariates or groupings of respondents. The analysis also seems unsatisfactory when we think of Example 2 with its 12 treatments and, hence, 66 pairs of treatments. However, a table like the one above but with 66 rows to describe performance of 12 varieties would be nothing but opaque!

The modeling approach to this type of data is based on the above table. The idea is to find a score  $s_i$  for each treatment such that the probability that treatment i is ranked higher than treatment j, when the two are compared, depends on the difference between the scores  $s_i - s_j$ . If the relationship between scores and probability is a logistic function, then the model can be fitted using standard logistic regression software. Hence:

$$p_{ij} = \text{Prob}(i \text{ ranked above } j) \text{ and}$$

$$\log(p_{ij}/1-p_{ij}) = s_i - s_j$$

Setting up the data to fit the model is slightly messy. There has to be a row for each pair of treatments compared. Thus, a farmer with just g and c will contribute 1 row of data for the pair g-c. A farmer with three treatments, g, s, and c, will contribute three rows of data, g-c, s-c, and g-s. Indicator variables are needed for each treatment, and the response variable contains 0s and 1s. The first few rows of data are shown:

Nome1	first1	second1	c	g	s	compl
Adisini	g	c	-1	1	0	1
Belo	s	c	-1	0	1	*
Belo	s	g	0	-1	1	1
Belo	g	c	-1	1	0	*
Bisiwiki	g	c	-1	1	0	1
Chakame	g	c	-1	1	0	1
Chakame	s	c	-1	0	1	1
Chakame	s	g	0	-1	1	0
Chiminbo	g	c	-1	1	0	*
Chiminbo	s	g	0	-1	1	*
Chiminbo	s	c	-1	0	1	*
Chinzeka	g	c	-1	1	0	1
Chinzeka	s	g	0	-1	1	0
Chinzeka	s	c	-1	1	0	

The first row of data shows that Adisini compared g and c. Treatment g was ranked higher than c, so when g is the first and c is the second, the response is “success”, indicated by a 1 in the last column. Belo compared all three treatments but the observation for c was missing, therefore both the s-c and g-c comparisons are also missing.

The modeling now proceeds similarly to that in other situations.

```
model [dist=b] compl; nbin=1
fit [con=o]g+s+c
```

```
*** Summary of analysis ***
```

	d.f.	deviance	deviance	mean deviance ratio
Regression	2	*	*	
Residual	74	73.49	0.9931	
Total	76	*	*	

\* MESSAGE: ratios are based on dispersion parameter with value 1

\*\*\* Estimates of parameters \*\*\*

	estimate	s.e.	t(*)	antilog of estimate
g	2.072	0.435	4.76	7.939
s	1.290	0.425	3.04	3.632

\* MESSAGE: s.e.s are based on dispersion parameter with value 1

The output looks a little odd because Genstat does not know what to use as a null model when the constant is omitted. It cannot calculate a total deviance and hence also cannot calculate a regression deviance. In this case the sensible null model is one of no preference between treatments, corresponding to  $p_{ij} = 0.5$  for all pairs, or  $\log(p_{ij}/1-p_{ij}) = 0$ . The deviance for this model is given by:

```
model [dist=b] compl; nbin=1
fit [con=o]
```

Now the analysis of deviance can be reconstructed:

	d.f.	deviance	deviance	mean deviance ratio
Regression	2	31.91	15.96	15.96
Residual	74	73.49	0.9931	
Total	76	105.4	1.386	

The model appears to explain much of the variation, suggesting real difference between the treatments. When interpreting the parameter estimates, remember that the  $p_{ij}$  value depends only on the differences  $s_i - s_j$ . Hence we only need to estimate two of the treatments and can arbitrarily set the third, in this case c, to zero. The estimates above therefore give an ordering and even magnitude of differences between the treatments. They can be compared with the results from analyzing both actual yields and the scores.

Treatment	Ranking		Rating		Yields		Scaled yields†	
	$s_i$	se‡	Log odds ratio	se	Adjusted mean	se	Adjusted mean	se
g	2.07	0.44	3.60	0.77	2.62	0.15	2.07	0.32
s	1.29	0.43	2.72	0.79	2.37	0.17	1.54	0.36
c	0.00	-	0.00	-	1.64	-	0.00	-

† Yield means are scaled to match the s scale of the ranking data.  
‡ se is the standard error of the difference from c.

The analysis of ranks has, to within the arbitrary scaling, produced an order and relative difference between treatments which is remarkably similar to that from the actual yield data, yet has larger sed values: the ranks contain less information than actual yields.

Note that the table of pairwise probabilities  $p_{ij}$  can be reconstructed from the scores  $s_i$  using the following relationship:

$$p_{ij} = \exp(s_i - s_j) / (1 + \exp(s_i - s_j))$$

These are shown in the table below and indicate a reasonable fit of the model.

Pair	No. of comparisons	No. with first of pair ranked higher than second	Proportion with first pair of ranked higher than second	Fitted probabilities $p_{ij}$
g-c	31	28	0.903	0.888
s-c	21	16	0.762	0.784
g-s	24	16	0.667	0.686

As in other situations, an advantage of using an explicit model to analyze the ranks, rather than relying on more ad hoc methods, is that the effects of covariates can be identified. To illustrate this I have looked at slope, classified into 2 levels (0 = flat, 1 = sloping), since one of the hypotheses was that g would perform relatively less well on sloping land.

add [p=a,e] slopel.(g+s+c)

```

*** Accumulated analysis of deviance ***

Change                mean deviance    deviance    ratio
   d.f.                deviance
- Constant
+ g
+ s
+ c                1                *
+ g.slopel
+ s.slopel
+ c.slopel        2                0.778        0.389        0.39
Residual          72                72.715       1.010
Total             75                *

* MESSAGE: ratios are based on dispersion
parameter with value 1
    
```

The analysis of deviance suggests that there is no consistent difference in the way g, s, and c are ranked on flat and sloping land. This conclusion is also reflected in the parameter estimates.

\*\*\* Estimates of parameters \*\*\*

	estimate	s.e.	t(*)	antilog of estimate
g	2.117	0.583	3.63	8.305
s	1.598	0.607	2.63	4.944
g.slopel 0	-0.056	0.901	-0.06	0.9454
g.slopel 1	0.000	*	*	1.000
s.slopel 0	-0.632	0.858	-0.74	0.5313
s.slopel 1	0.000	*	*	1.000

\* MESSAGE: s.e.s are based on dispersion parameter with value 1

These values can be combined in a table of scores, together with standard errors of the difference between treatments within slope categories.

Treatment	Slope = 0	Slope = 1
g	2.117 - 0.056 = 2.061	2.117
s	1.598 - 0.632 = 0.966	1.598
c	0	0.000

If the standard errors of the interaction effects were smaller, we would say the results were consistent with the hypothesis—that the difference between g and s is greater on flat than on sloping land.

Remember it is impossible to look at the main effect of slope. We cannot determine whether the treatments are generally assessed as better on flat than sloping land. Each participant ranks among the alternatives tested on their farm, and each farm is classed as either sloping or flat. Similarly we cannot compare the two columns in the table above, i.e., compare g on flat and sloping land. There is no information in the data

on this comparison, as all rankings are done within farms. The situation would be different if there were farms that had both flat and sloping land.

**Example 2**

Yields for Example 2 were also converted to ranks for the purpose of illustrating the analysis. Remember this study comprises 12 varieties, with 146 farmers each comparing 4 varieties. It is difficult to think of a useful, simple, descriptive analysis of this rank data that shows the differences between varieties. The design is very unbalanced, so any simple totaling of ranks will give a biased picture. We could look at all  $12 \times 11/2 = 66$  pairwise comparisons, and find the proportion in which one treatment ranks above another; however, it is not easy to view a matrix of 66 values and understand the relative performance of 12 varieties. It has been suggested by Russell (1997) that an overall score be given to each variety by counting the number of times each one ranks above another; however, this requires that each occurs equally often. Some sort of average proportion could be devised; however, the modeling approach is simple once the data file is set up.

The data file structure and modeling proceed as per Example 1. Twelve indicator variables,  $e[1], \dots, e[12]$ , are needed for the 12 varieties. In the statements below the first FIT gives the correct total deviance from which the analysis of deviance table is constructed.

```
model [dist=b] compl; nbin=1
fit [con=0]
fit [con=0] e[1...12]
```

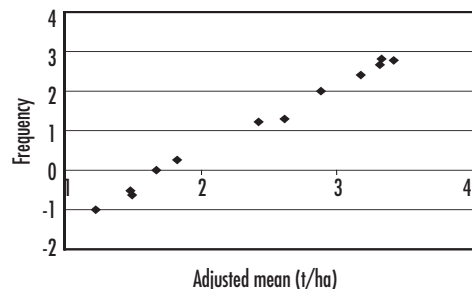
	d.f.	deviance	mean deviance	deviance ratio
Regression	11	439.1	39.9200	39.92
Residual	865	774.9	0.8959	
Total	876	1214.0	1.386	

\*\*\* Estimates of parameters \*\*\*

	estimate	s.e	t(*)	antilog of estimate
e[1]	-0.996	0.293	-3.40	0.3692
e[2]	1.998	0.300	6.66	7.378
e[3]	1.282	0.284	4.52	3.606
e[4]	2.818	0.317	8.88	16.74
e[5]	-0.523	0.303	-1.73	0.5926
e[6]	2.655	0.312	8.51	14.22
e[7]	0.247	0.277	0.89	1.281
e[8]	1.205	0.281	4.29	3.336
e[9]	-0.637	0.299	-2.13	0.5289
e[10]	2.769	0.318	8.70	15.95
e[11]	2.423	0.303	8.00	11.28

\* MESSAGE: s.e.s are based on dispersion parameter with value 1

Genstat has put the score for the last treatment,  $s_{12}$ , to zero. The parameter estimates above give scores that show the relative performance for each variety. If these are compared with the results based on actual yields, it can be seen (in the graph below) that the method not only reproduces the ordering of the varieties very closely, but also the relative differences. Of course in this case the ranks were calculated from the yields without error. However, it still seems very surprising that this information about relative performance of the varieties can be recovered from just the four ranks on each farm.



As each score is relative to the score of zero for variety 12, the se values listed with the estimates are for the comparison of that variety with variety 12. Other se values are most easily found using `predict`. For example, the difference between scores for variety 1 and 2 is found by:

```
predict [back=n] e[1...11]; 1,-1,0,0,0,0,
0,0,0,0,0
```

```
Prediction S.e.
-2.995      0.335
```

More complex contrasts between treatments can be similarly calculated. For example, varieties 1, 5, 7, 9, and 12 are in one group, a. Varieties 4, 10, and 11 form group b. We can calculate the difference between the average scores for groups a and b by taking  $(s_1+s_5+s_7+s_9+s_{12})/5 - (s_4+s_{10}+s_{11})/3$ . Remembering that  $s_{12} = 0$ , `predict` can be used for this:

```
predict [back=n] e[1...11]; 0.2,0,0,-
0.3333,0.2,0,0.2,0,0.2,-0.3333,-0.3333
```

```
Prediction S.e.
-3.052      0.213
```

Group a is clearly worse than group b.

As in Example 1, it is simple to turn differences in scores into probabilities of one variety being ranked higher than another. For example, the chance that variety 1 is ranked higher than 2 is given by:

```
predict e[1...11]; 1,-1,0,0,0,0,0,0,0,0,0
```

```
Prediction S.e.
0.0477      0.0152
```

Variety 2 is almost certain to be ranked higher than variety 1.

As before, the model can now be extended to look at the extent to which covariates interact with treatment differences. I use two continuous covariates: soil P and sand content. The data file has been set up with a column giving the sand and P value for each pairwise comparison.

```
fit [p=*; con=o] e[1...11]
add [p=*; con=o] sandfl.e[1...11]
add [p=a; con=o] Pfl.e[1...11]
```

The analysis of deviance table can be constructed from this output. Note the total degrees of freedom has changed from that given earlier as there are missing values in the covariates.

		d.f.	deviance	mean
e[1...11]		11	309.50	28.140
+sandfl.e[1...11]		11	15.64	1.420
+Pfl.e[1...11]		11	15.16	1.380
Residual		561	483.30	0.8614
Total		594	823.50	1.386

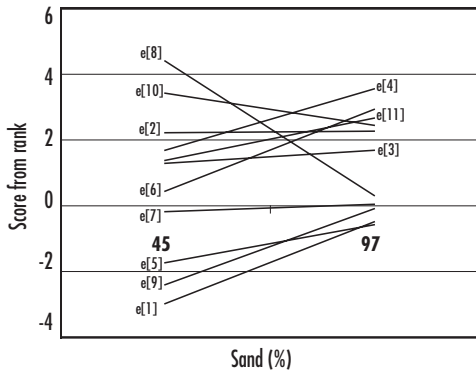
The results show that neither P nor sand have a strong interaction with variety. However in order to show the types of results obtainable, the model with sand is refitted and parameter estimates produced.

```
fit [con=o;p=e] e[1...11]+sandfl.e[1...11]
```

\*\*\* Estimates of parameters \*\*\*

	estimate	s.e.	t(*)	antilog of estimate
e[1]	-5.11	2.83	-1.81	0.006012
e[2]	2.22	3.21	0.69	9.223
e[3]	0.97	2.49	0.39	2.631
e[4]	0.10	3.20	0.03	1.109
e[5]	-2.73	2.61	-1.04	0.06530
e[6]	-1.72	2.75	-0.62	0.1796
e[7]	-0.42	2.38	-0.18	0.6559
e[8]	7.90	3.70	2.14	2699.
e[9]	-4.38	2.64	-1.66	0.01248
e[10]	4.28	3.46	1.24	72.58
e[11]	0.28	4.01	0.07	1.317
e[1].sandf1	0.0475	0.0327	1.45	1.049
e[2].sandf1	0.0004	0.0370	0.01	1.000
e[3].sandf1	0.0075	0.0290	0.26	1.008
e[4].sandf1	0.0357	0.0375	0.95	1.036
e[5].sandf1	0.0223	0.0301	0.74	1.023
e[6].sandf1	0.0478	0.0320	1.49	1.049
e[7].sandf1	0.0049	0.0280	0.17	1.005
e[8].sandf1	-0.0781	0.0417	-1.87	0.9249
e[9].sandf1	0.0444	0.0304	1.46	1.045
e[10].sandf1	-0.0187	0.0398	-0.47	0.9814
e[11].sandf1	0.0244	0.0460	0.53	1.025

The scores for each variety now depend on the sand content, for example, the score for variety 1 is  $s_1 = -5.11 + 0.0475$  sand. These are plotted below for the range of sand contents (45-97%) found in the trial.



Remembering that the scores show the relative performance of varieties, with variety 12 fixed at a score of zero, two main patterns emerge. Several varieties (1, 4, 5, 6, and 9) with increasing sand content rank higher than variety 12. Variety 8 with high sand content ranks distinctly worse.

## Discussion

### General

The methods for analyzing ranking and rating data described above are not new, but they have not been routinely used in the analysis of agricultural trials. A discussion of the proportional odds model used for rating data can be found in Agresti (1996). The model for ranks is not so widely used, explaining why common statistical software does not make it immediately available. When the observations are just paired comparisons (i.e., each participant is asked to state which of two treatments is superior), the model used is the Bradley-Terry model (Bradley and Terry 1952), which has been widely used, particularly in social science applications. Dittrich et al. (1998) use the method for paired comparisons when there are categorical covariates, and mention that it is possible with continuous covariates. The approach used when more than two treatments are compared is described by Critchlow and Fliener (1991).

Both models make assumptions about the nature of the data; however, this is true of all statistical analyses. It is a necessary part of attempting to reach conclusions about general patterns. Methods for checking the key assumptions are well developed for

established linear model methods (e.g., looking for various patterns in residuals), and similar tools need developing for the models discussed. Alternative models may be more appropriate for ranks and rates. The methods presented here appear to be the simplest of those that have proved useful in some common situations. Again this is true of all statistical modeling. For example, linear regression analysis is widely used but not because “nature has to be like that” but because the model has been found to be a useful approximation for many problems.

From the examples in this paper it should be clear that statistical analysis of participatory breeding trials cannot be automatic. When researcher designed trials were run using a very regular design, it was possible (though probably not wise) to run a standard analysis on each data set. Such an approach will not recover most of the useful information from participatory trials.

### Further discussion on analysis of ranks

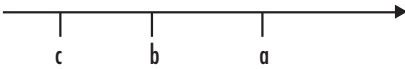
The method described above for analyzing data presented in the form of ranks is appealing and powerful. It is able to produce an overall ordering of treatments and even indicate the relative magnitudes of the differences between treatments. It can handle awkward incomplete sets of data in which each farmer does not rank all treatments. Most importantly it can show how treatment order interacts with covariates. In Example 1, the covariate was a categorical variable, dividing the sample of farmers into groups. Continuous covariates were analyzed in Example 2.

Unlike many other approaches to analysis of ranks, the method uses estimation, not just testing. This distinction is often made when analyzing continuous variates such as crop yield. It is rarely beneficial to simply conclude that mean yields differ significantly between varieties, or even that variety A yields significantly higher than variety B. Useful conclusions can be drawn when we are able to identify by how much A outyields B, and put a confidence interval around this. The same is true when analyzing ranks. It is rarely going to be useful to simply report that treatments differ significantly in their ranks, yet this is all that most statistical procedures for analysis of ranks do. The method presented here shows the relative magnitude of the differences, and these can be interpreted. For example, we may show that A and B are ranked significantly differently. The scores for the varieties can be converted to a probability  $p_{AB}$  that A will be ranked higher than B. If  $p_{AB} = 0.95$ , the interpretation is very different than if  $p_{AB} = 0.55$ , yet both could be significantly different from the no-preference value of  $p_{AB} = 0.5$ .

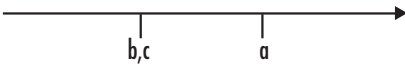
There are a number of questions regarding this analysis, some of which require theoretical statistical investigation.

1. The model makes assumptions about the nature of the data and the effects of treatments and covariates. It is not clear how to check whether they are reasonable or how robust the results are to departures from the assumptions.

2. The analysis depends on the model which assumes the treatments can be allocated scores such that the probability of one ranking higher than another depends on the difference between the scores. This is the “linearity assumption” made by Taplin (1997). It is helpful to represent this graphically. If farmers consistently rank  $a > b > c$  then we could derive scores that would put the treatments on a line:

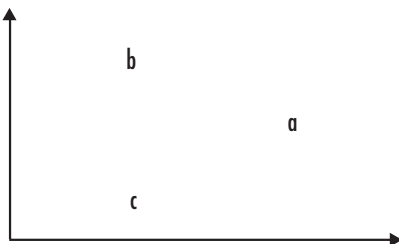


It is easy to produce data for which this linearity assumption fails. For example, we can have  $a > b$  and  $a > c$  equally often, suggesting the ordering should look like the line below.



However at the same time we can have  $b > c$ . This might occur if, for example, different farmers were making the comparisons and using different criteria for each one.

The problem occurs in other examples of ordination, for example, that used by ecologists to describe species occurrence. One answer is to introduce a second dimension. The distances between point a, b, and c can reflect the rankings if they are arranged as follows:



With just three treatments (as in Example 1) it is clear how an extra term can be introduced in the model to test whether a non-linear arrangement is superior. (In Example 1 it is not.) However, I do not know how to fit models that follow the usual, useful multivariate approach of gradually increasing the number of dimensions until a suitable fit is obtained.

3. In some ranking procedures ties are allowed, for example, farmers are allowed to state that they have no preference between two or more of the treatments. Dittricht et al. (1998) show how the model can be modified to allow for ties. Extra parameters are included so that for each pair we estimate the probability that they tie, as well as the probability that one is ranked above the other.
4. Coe (1, this proceedings) illustrates the value of being able to describe variation at different levels in the design with random effects. It is not clear if these ideas are useful or could be used here. In principle we can fit the model with random effects using the GLM framework; however, maybe it is not necessary. All of the information in the ranks is at the within farm level; hence, we can look at treatment differences and interactions of these within-farm level (or higher level) covariates. However we cannot look at any between farm effects. It is also not clear if plot level covariates can be incorporated. For example, suppose farmers ranked treatments but also reported whether each plot was normally fertile or not, so that plots within one farm could have differing values of the fertility covariate. A model that uses this data would have to be built on the probability of  $a > b$ , depending on both the difference in treatment scores and the difference in fertility.

5. The analysis described here is suitable for one objective—for determining treatment effects and their interaction with covariates when the observations are ranks and the design is incomplete or irregular. If the design is more complete, for example, with each respondent comparing all treatments, then other approaches are possible. Different objectives may be of interest, for example, comparison of the rankings under different criteria, or partitioning the sample of respondents into homogeneous groups, when again different methods will be appropriate. Remember also that if the data are rankings produced at, say, a group meeting, so that a consensus is achieved, then no statistical analysis is necessary. Abeysekera (2001) and Riley and Fielding (2001) describe some of the simple alternatives. Taplin (1997) describes a number of the statistical tests available.

### Ranks or rates

Having seen how data from these trials can be analyzed, it is worth looking again at the relative merits of using ranking or rating.

First it should be clear that a response measured on a continuous scale, using an accurate and unbiased instrument, contains more information than the equivalent observation using a rating scale with a few levels or than using ranks. Reasons for not using the continuous variate include:

1. Time, money, and logistics, e.g., we may not be able to measure crop yield because we don't know when farmers will be ready to harvest.
2. Lack of a suitable instrument. If we want to assess taste or opinions, there is not much alternative to rating or ranking.

3. Participation. Collecting ranking and rating data requires participants. Other measurement methods may be alienating.

Methods of collecting rating data have been described (e.g., Ashby 1990) and include tools that can give high quality, repeatable and reliable data. It appears that farmers are able to give scores to a large number of alternatives. There are statistical questions regarding the number of levels to use. There is no point in using too many levels, as small differences in rating will probably not reflect real differences in opinion. Note that we do not make a rating scale into a continuous variate simply by using many levels. The fundamental characteristic of a rating scale is that the numbers represent qualitative labels (very good, poor, etc) and the quantitative analogue is missing. This may not be the case if markers are used to represent the score. There is a lot of theory and practice from the social science literature that is relevant here. Respondents are often reluctant to use the ends of a scale, particularly the lower end. Hence a 5-point scale may in practice be used as a 3-point scale. Note that some degree of consistency in the use of the scale by different participants (particularly in different locations) can be achieved by explaining what poor, excellent, etc, mean. For example, poor might correspond to "I would never consider growing this again", while excellent might mean "I would like this to become a main variety on my farm each year".

Ranking is used when it is thought that participants might find it easier to order alternatives than to give them a score.

One reason for this is clear: participants may have a preference for two alternatives given the same score (e.g., both excellent), and hence be able to give them different ranks. However, a shortcoming is immediately clear: the ranks may be the same if both alternatives are also considered poor. This is an important problem. The information in ranks is all “within respondent”, that is, we can identify whether, for example, participants consistently rank A above B, and we can determine whether this is true for both male and female respondents. However, we cannot determine what either group of participants actually thinks of A and B. An important part of any research is to make generalizations and extrapolations; however, this often not possible from rank data. Abeysekera (2001) makes the point that the information in ranks is enhanced considerably if some sort of baseline is also measured. For example, if a local control variety is included in each participant’s set of alternatives, then we could find a rating for the local control and rank the others relative to this. It is not clear exactly how such data could be analyzed.

A study by ICRAF (1996: 55) assessed the suitability of 12 tree species for firewood. The researcher thought that women could only realistically compare pairs of species. The participants ranked each pair tested, from which it was possible to produce an overall ordination. However, they were also asked the reasons for preferring one species to the other. An alternative design would have used a pilot study of this type to elicit important criteria and then asked these to be rated for each species tested.

Remember there is no information in ranks on effects of quantities that vary across farms. In Example 1, we were unable to determine whether g was more effective on sloping or flat land.

Overall there seems to be little reason to use rank data unless they are specifically required according to the objectives.

## References

- Abeysekera, S. 2001. *Analysis approaches in participatory work involving ranks or scores*. DFID Theme Paper (revised). UK: Statistical Services Centre, University of Reading.
- Agresti, A. 1996. *An introduction to categorical data analysis*. New York: Wiley.
- Ashby, J.A. 1990. *Evaluating technology with farmers: a handbook*. Cali: CIAT.
- Bradley, R.A., and M.E. Terry. 1952. Rank analysis of incomplete block designs, 1. The method of paired comparisons. *Biometrika* 63:255-262.
- Coe, R., and S. Franzel. 2000. Keeping research in participatory research. Presented at the 3<sup>rd</sup> International PRGA Conference, 6-11 November, Nairobi, Kenya.
- Coe, R., R. Stern, and E. Allen. 2001. *Analysis of data from experiments*. N.p.: SSC/ICRAF. 250 pp.
- Critchlow, D.E., and M.A. Fligner. 1991. Paired comparisons, triple comparisons and ranking experiments as generalised linear models, and their implementation in GLIM. *Psychometrika* 56:517-533.
- Dittrich, R., R. Hatzinger, and W. Katzenbeisser. 1998. Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Applied Statistics* 47:511-525.
- Genstat. 2000. *Genstat for Windows*. 5<sup>th</sup> Ed. Oxford: VSN International.
- International Centre for Research in Agroforestry (ICRAF). 1996. *Annual Report 1995*. Nairobi: ICRAF.
- Riley, J., and W.J. Fielding. 2001. An illustrated review of some farmer participatory research techniques. *Journal of Agricultural, Biological and Environmental Statistics* 6:5-18.
- Russell, T. 1997. Pair wise ranking made easy. *PLA Notes* 28:25-26.
- Taplin, R.H. 1997. The statistical analysis of preference data. *Applied Statistics* 46:493-512.

## Discussion Summary

The discussion centered on three themes: (1) the advantages and disadvantages of ranking versus rating, (2) the relationship between rating and parametric and non-parametric tests, and (3) the use of ranking in the context of group interviews, particularly the popular participatory technique of matrix ranking. In terms of the pros and cons of ranking versus ratings, ranking imposes a more rigid framework than rating and ties are not allowed. It is difficult to rank more than six varieties. For example, in an initial mother trial of 52 varieties it is impossible to rank them all, but in a baby trial with only 4 to 8 varieties, ranking is feasible. It was pointed out that with many varieties, farmers may be able to rank the first few, but the rest probably would be considered equally (poor) and would not be ranked effectively. Another problem with ranking is that it may not be possible to know how good or poor a variety really is. Varieties may be ranked, but may all still be considered poor, or vice versa. Different groups of farmers may rank and score differently. For example, ranks for cassava varieties change for different characters. It was pointed out that another potential disadvantage of the use of ratings is that they can vary from village to village; however, ratings are appropriate where a farmer is comparing a new variety against a control in a binary fashion such as “better” or “worse”. It was also pointed out that the analysis of this type of data is not obvious and merits further research. A question posed was whether pair-wise comparisons (rating/ranking) can be made within a group of varieties. The answer was that the data would probably be more inconsistent than ranking/rating all varieties. Another important question was whether there were trade-offs between the use of ratings versus rankings. This is an aspect that has not been widely considered.

Another question was related to the appropriateness of non-parametric tests for rating/ranking. The answer was that such tests do not work for unbalanced designs; they are only testing procedures and do not allow an estimation of the effects of factors on the response. They are therefore less powerful than the methods described in the lecture. Furthermore, these procedures can be extended to genotype x environment analysis. It was pointed out that if ratings are assumed to be normally distributed, then a probability approach could be used. (Editor’s note: I assume this point refers to the use of parametric procedures). This would depend on whether the researcher is happy with the assumption that there is a continuous variable underlying the ratings, and that, in fact, the mean of a 3 rating and a 5 rating is 4.

The final discussion covered matrix ranking. This is a popular participatory technique used with farmer groups for assessing varieties. Group discussion and consensus are often very useful to begin with, particularly for an initial synthesis or screening inventory, but cannot provide an analysis of the heterogeneity present. In some cases consensus may be imposed by the structure of the exercise. Consensus is not necessarily of great interest as often there is more benefit in examining heterogeneity. The value of matrix ranking is limited because it is carried out with groups where consensus is sought, and the results cannot be analyzed statistically since there is no variation (few degrees of freedom). A question was asked on whether the same people are used for ranking the different characteristics. The answer was that usually this is not the case—different people rank different characteristics, compounding the difficulty of analyzing the results.