

# The Application of Rasch Scaling to Wine Judging

Murray Thompson

Flinders University School of Education [dtmt@senet.com.au](mailto:dtmt@senet.com.au)

*The training of judges in sport and in industry is a challenging educational problem and recent developments in educational measurement can contribute to the resolution of this problem. The results of the judging of 98 wines from one class of a prominent Australian wine show were analysed using Rasch scaling. The raw scores were out of 20, graduated in steps of half a point, giving a total of 41 different scores. It was necessary to reduce this scale to an eight point scale, making all scores below 12.5 zero and compressing the remaining scores into seven scores from one to seven with each score combining two of the raw scores. It was found that this compression process does not adversely affect the analysis. This suggests that the traditional judging process tries to apply too fine a scoring system.*

*Two programs were used to estimate the harshness of the judges and the level of the wines. They were the Quest program and the RUMM program. It was found that there was a good fit to the Rasch model and useful, but some conceptually different information was derived from both the Quest and the RUMM programs. It was therefore possible to put the quality levels of the wines and the harshness of the judges on the one scale and make useful observations about the harshness and consistency of each of the judges which should prove useful feedback for the training and on-going professional development of wine judges. The RUMM program in particular offers very useful feedback to judges that shows just how they are awarding their grades along the continuum.*

*A means of reducing the total number of wines to be tasted by each judge was explored and it was found that the Rasch scaling procedure could be used to reduce the total tasting load on judges. Comparison of the raw scores the Quest and the RUMM showed very high correlations, with evidence of a greater spread of the Quest scores than the RUMM scores.*

Wine judging, Rasch scaling, Quest, RUMM, consistency

## INTRODUCTION

Imagine a prestigious scholarship or prize to be judged on the quality of a single essay from several hundred candidates by a panel of judges, each assessing the performance of each candidate. What procedures could be put in place to ensure that the most worthy candidate wins the prize? Picture an Olympic Games Diving final, with a diver poised on the 10m tower and a line of judges ready to assess her performance for a possible gold medal. What safeguards can be put in place to ensure that the judging process is fair and not subject to bias of the judges? A row of 200 red wines, lagers, cheeses or olive oils stands awaiting the judgment of a panel of judges. How can the judging panel be certain that their decisions really reflect the quality of the products?

All of these situations have a number of things in common. They involve the complex process of the judgement of a single performance of each of the subjects, assessed by a number of raters. They represent situations of significant advantage to the successful candidate or producer of the

product and it behoves the judges and the organizations that they represent to make the judging process as fair, as accountable and as transparent as possible.

Thus, it is clear that the training of judges presents a challenging problem for educators that lies outside the traditional fields of primary, secondary and tertiary education. Moreover, both sport and industry demand that their judging processes should be of the highest quality, giving considerable time and effort to the training of judges, but without using the knowledge of measurement that has been developed by research workers in education during the past two decades to deal with the judging problem. This article is written for publication to advance the argument that educators should look beyond their traditional fields of pedagogy and research and recognize that educational processes are involved in many fields of endeavour outside the organizations of primary, secondary and tertiary education. The processes of education operate wherever people are learning and teaching and at every stage in their lives, and educators and educational research workers should accept the challenge of contributing their particular knowledge and expertise to the resolution of common problems.

Cronbach (1964, pp.506-511) outlined some of the problems of making judgements. Referring particularly to supervisors making ratings of their subordinates, he outlined a number of sources of error in the judgement process. The first is 'generosity error' where the rater gives very favourable reports in all but the very worst cases and therefore the reports do not discriminate well. A second problem he cited was 'ambiguity' which is associated with varying interpretations of the criteria. A third source of error that was identified was 'constant error' or 'bias.' For example, one rater may not use the extremes of a scale while another may do so. A further concern is related to 'limited information' which may be available about the individual being judged and the final problem is the 'halo effect' in which an overall opinion might obscure some serious undesirable traits. He suggested that the reliability of ratings can be improved by combining the ratings of several judges. An alternative is to keep records of the ratings of a particular judge to establish and therefore correct any errors. The use of 5 to 7-point scales was suggested as a useful strategy that directs the rater to the kinds of deviation being explored.

Wolf (1997) discussed the use of rating scales, suggesting a number of ways in which their use can be improved. These included the use of multiple raters, and the training of raters. The use of Item Response Theory (IRT) methods was also suggested as a way of assigning the values of the scale.

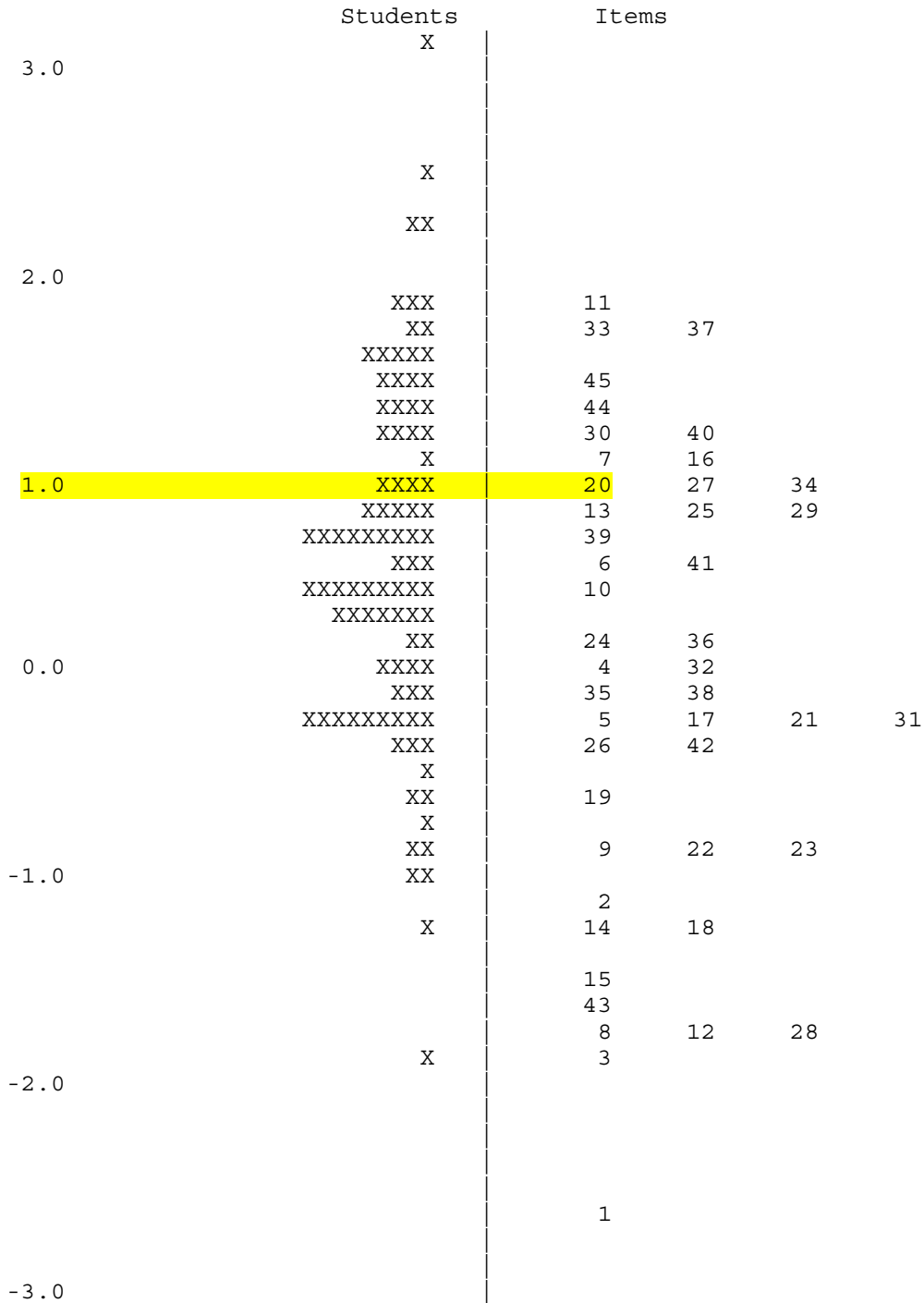
Over the years, there have been developed a number of systems of checks in the judging processes used to rate various products. One particularly common practice is to include several samples more than once to test the consistency of the judgement at various times during the proceedings. However, there is no system in place to allow, either for the huge numbers of tastings or assessments required or the other problems of integrity involved. In diving, the highest and lowest scores are removed from the final score to reduce the spread of scores in order to eliminate outlier judgements.

### **Rasch Scaling**

Rasch Scaling has been used in educational measurement to overcome similar difficulties to those described above. Bond and Fox (2001) have outlined the use of Rasch model in a range of measurement problems. Central to the Rasch model is the employment of a unidimensional scale that is used to define both the performance of the students taking a test and the difficulty of the items or questions in the test. This scale is graded in logits.

In Figure 1, the logit scale is shown on the left, ranging from  $-3.0$  to  $3.0$ . This scale is commonly calibrated to make the average difficulty of the items 0. The performance of the students taking

the test is represented by an x and the difficulty of each of the numbered questions or items is shown on the right. In the diagram, Item 20 has a difficulty of 1.00. This means that a student whose ability is also rated at 1.00 has an equal chance of getting Item 20 right or wrong. These difficulty levels are referred to as Thurstone threshold values and involve a 50 per cent probability of attaining a score on either side of the threshold value.



**Figure 1. The performance of a group of students plotted on the same scale as the difficulty of the items or questions in the test**

These ideas have been extended to incorporate more complex testing procedures, including allowing for various levels of performance in each question. Thus each item may have several

grades attached to it and students may be awarded a score in the range 0-7 for their performance on each question. This is known as the partial credit model and the difficulty of each item is subdivided into several levels. Thus for a particular item it is possible to describe the difficulty in the following manner. If an item has eight levels, 0-7, then a threshold between two of the levels is set in calibration as that scale distance between two grade scores. Thus, on a 0-7 scale, Threshold 4 would be the level at which a score of either 3 or 4 is equally likely to be awarded. Above this threshold level a 4 is more likely and below this threshold a 3 is more likely.

Once a set of test scores has been Rasch scaled, it becomes possible to use the scales to make important comparisons and connections. For example, it is possible to equate two different tests to one another and therefore make meaningful comparisons between students who take different tests. Similarly, it is possible to make comparisons between markers of essay questions so that essay scripts can be marked by a team of markers, without the need for every member of the team to mark every script. One of the important issues that can be explored using Rasch scaling is the question of judge bias.

### ***Rasch scaling and judging***

Recent advances in educational measurement techniques have allowed some very worthwhile analysis of testing. Of particular interest is the application of Rasch scaling to the area of judgment. This has been used to examine the harshness of essay markers in order to standardise the marking procedures as well as to determine the level of difficulty of questions and the abilities of the students.

Linacre discussed the use of Rasch models in the measurements of judgements. "A Rasch model analysis capitalises on the inevitable judge disagreement to construct a sample-free, objective and linear measurement continuum." (Linacre 1999, p 244) He suggested that in a judging system involving the Rasch model, the only requirement is that there must be a connected network of each rater, each candidate and each assessment item. (Linacre 1999)

Andrich (1999) has applied such analysis to equate the marks given in essay questions. In this study the effect of question choice on the final mark was explored and a means of equating the marks in several different questions was developed. Thus the analysis was used to compare the marks of students who had selected one question that was seen to be more difficult than another. The point was made that with a minor variation in design, it would be possible to explore the effect of the raters, rather than the difficulty of the question.

Bond and Fox (2001) discussed judged sporting performances, citing particularly the work of Looney (1997) who used Rasch scaling procedures to establish Eastern Block bias in the Free Skate event the 1994 Winter Olympics.

The purpose of this analysis is to investigate the application of educational measurement techniques to some of the problems and challenges associated with the judging of wine. These problems include:

- variable judge harshness or difficulty,
- judge consistency,
- issues associated with the large numbers of entries,
- training of judges and giving appropriate feedback.

The data that have been used are from a leading wine show and a particular focus has been given to a particular class with a large numbers of entries. This was the cabernet sauvignon class. Each exhibit has been judged by six judges. These judges include three expert judges and three

associate judges. Each judge rated the wines out of a total of 20 points with steps of  $\frac{1}{2}$  point. Thus there were 41 possible scores that could be awarded by the judges. In practice, the range that was used by the judges was very much less than this. In the case of the cabernet sauvignon class, there were 102 exhibits, with four being withdrawn, leaving 98 wines that were judged in this class. From these wines, the lowest score was 10 and the highest was 19. It follows that in practice there were only 19 scores that were used by the judges. As well, it was found that scores at the extremes of these margins were rarely used, if at all by some individual judges.

For the award of medals, the scores of the three expert judges were totalled and so each wine was awarded a score out of a maximum of 60 points. It was assumed that the associate judges were being trained and that their scores were not included in the final tally. Thus, in the competition, there was no allowance made for the difference in the harshness of the various judges and that all of the judges were assumed to rate equally. However, it must be recognized that the judges had been trained over a long period, and consequently they might be expected to have standards that were highly similar in the assessment of wines of different types.

## **Problems with the Traditional System**

### ***Leniency and harshness***

In the traditional system no account is taken of the leniency of the judges. For example, one judge may simply award all wines a point higher than the others. Similarly, another judge may be more harsh and so award all wines points lower than the other judges. This in itself may be seen to even out but for the sake of consistency between judges, it is desirable for judges to be given clear feedback about how they are awarding points compared with one another on a clearly defined and easy to interpret scale. The final points then can be weighted to take into account the variation in judge harshness.

### ***Consistency***

The question of judge consistency is also a difficult one. As described above, allegations have been made which suggest that in some competitions judges associated with particular business interests recognise their own products and award them high points and downgrade the points of direct competitors. (White 2000) The traditional system has no mechanism for examining the level of consistency of the individual judges. Quite apart from the concern of deliberately rating a particular product up or down, individual judges would be greatly assured by evidence of their consistency. It would be in effect a form of quality control for their work. This problem becomes worse as judges are required to judge more and more wines. The confusion and tiredness of the palate must become a concern. With the current analysis of 98 wines, the levels of concentration and memory required combined with the confusion of flavours must make this a very demanding task. Even the most experienced judge of a high level of integrity must have concerns about coping with this onerous responsibility.

## **THE APPLICATION OF RASCH SCALING TO WINE JUDGING**

Of interest to wine judges is the question of how they award their scores for wines compared with how their colleagues do. Also of interest is how consistent is each judge. Under the proposed model, the judges take a role equivalent to items or questions and the wines have a role equivalent to the students. The analysis seeks to examine the level of each of the thresholds for each of the judges and the performances of each of the wines, placed on the same scale for easy comparison. The 41 point scale used by judges presents many problems, particularly as most of these scores are not typically used by the judges and, even within the usable range, scores on the extremes are

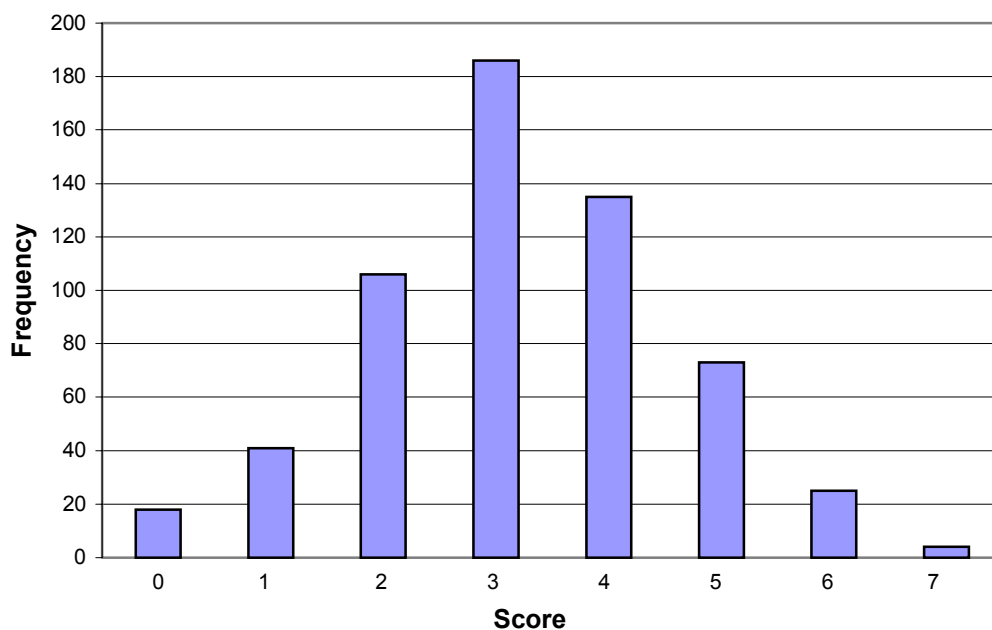
rarely used. Moreover, the computer programs require that a sufficient number of wines are associated with each score level for effective calibration and estimation of the threshold levels. Accordingly, the scores were compressed according to the following scale into an eight point scale prior to analysis.

**Table 1. Conversion of the raw scores to an 8 point scale**

Score awarded by judge /20	Converted to 8 point scale	Frequency (Total 588 = 6 x 98)
≤ 12.5	0	18
13, 13.5	1	41
14, 14.5	2	106
15, 15.5	3	186
16, 16.5	4	135
17, 17.5	5	73
18, 18.5	6	25
19, 19.5	7	4

It is readily seen that this scale compresses the scores at the lower end but retains most of the sensitivity at the upper end where the scores tend to be crowded and where greater discrimination is required. These transformed scores can then be analysed using two Rasch analysis programs. The two programs employed are the Quest (Adams and Khoo, 1993) and RUMM (Andrich et al. 2000) computer programs that tackle the problem of estimation of parameters in different ways. Moreover, these two programs present the user with somewhat different information on the results of analysis. Each supplies different information on the scaling process and these presentations are largely complementary in nature.

Figure 2 shows the frequency histogram of the scores after they had been transformed to the 0-7 scale. Of particular importance is the bell-shaped curve reflecting an essentially normal distribution of the scores.



**Figure 2. Frequency histogram of cabernet sauvignon data**

The initial focus was on the cabernet sauvignon class, simply because this was the class with the largest number of entries. The first program to be used was the Quest program that estimated the performance of the wines and the harshness of the judges using a joint marginal maximum

likelihood estimation procedure, with the partial credit model. (Adams and Khoo 1993) This procedure requires a generally normal distribution of scores, which is true in this case, as illustrated in Figure 2. This program can be applied to a scale calibrated in a situation such as is presented by the judging of a class of wines. This analysis can be carried out in a number of ways so that its application can be explored.

The first analysis was to examine the scale calibration with all six judges, both experts and associates. In this analysis, Item 1 was Judge A1, Item 2 was Judge A2 and Item 3 was Judge A3. These were the associate judges. Item 4 was Judge C, Item 5, Judge B and Item 6 was Judge P.

In the situation of calibrating a scale for a group of wines, the wines are the students and the items become the judges. It thus becomes possible to rate the performance of each wine, the harshness of each of the judges by identifying where on the scale their scores lie and the consistency of the judges in scaling on a unidimensional scale.

### Scale calibration

The Cabernet Sauvignon data were first analysed using the Quest analysis program using a partial credit model. Figure 3 shows the single scale used to rate both the performance of the wines themselves and the harshness of the judges. On the left, each of the wines is represented by  $x$  and it can be seen that the display indicates a well-shaped bell-like curve as would be expected with a normal distribution. On the right are displayed the thresholds in the scale levels for the various judges and between the eight different score levels. These thresholds are the points along the difficulty scale at which the individual judge selects the next highest score. For example, 1.1 is the level above which Judge 1 would award 1 and not 0. Similarly 6.4 is the level at which Judge 6 would be equally likely to award 4 or a 3. Each threshold level is then the changeover scale level for each of the boundaries between the eight possible scores. These have been shaded to allow for the ready identification of individual judges. Note the differences in scale distance between threshold levels for different judges. As well as the individual thresholds on the right of the diagram the overall harshness of each of the judges is plotted.

An estimate of the consistency of each of the judges can be obtained using the infit mean square diagram, shown in Figure 4. In this diagram, it is desirable for the values estimating the degree of fit of the judges to the unidimensional scale to be within the dashed lines on either side. The ideal is 1. Deviation to the right indicates that the particular judge has been inconsistent, perhaps by awarding a wine with a low score when the other judges have awarded a high score. It can readily be seen that Judge A1 has been the most inconsistent, although this judge is within the acceptable limits indicated by the dotted vertical line. Deviation to the left indicates a high degree of consistency. This is particularly the case for Judge B, who is so consistent with the rest of the group nothing much new is being added by his opinion. What is very clear here is that the experienced judges C, B and P all show very strong consistency, while the associate judges A1, A2 and A3 are tending towards a lesser degree of consistency. This would be very useful feedback for inexperienced judges. The actual values of the infit mean squares for each judge are given, along with the discrimination indices for each judge shown in Table 2. The judges are considered to be items in normal test and item analyses.

In order to summarise these results, it is noted that the judging process does fit the Rasch model well and that the wines and the judge harshness can be placed on a single scale which enables estimates to be made of the performance of each of the wines and the harshness of each of the judges. Further, it is possible to examine the consistency of each of the judges. What is particularly noticeable is the consistency of the experienced judges compared with the associate judges in this analysis.

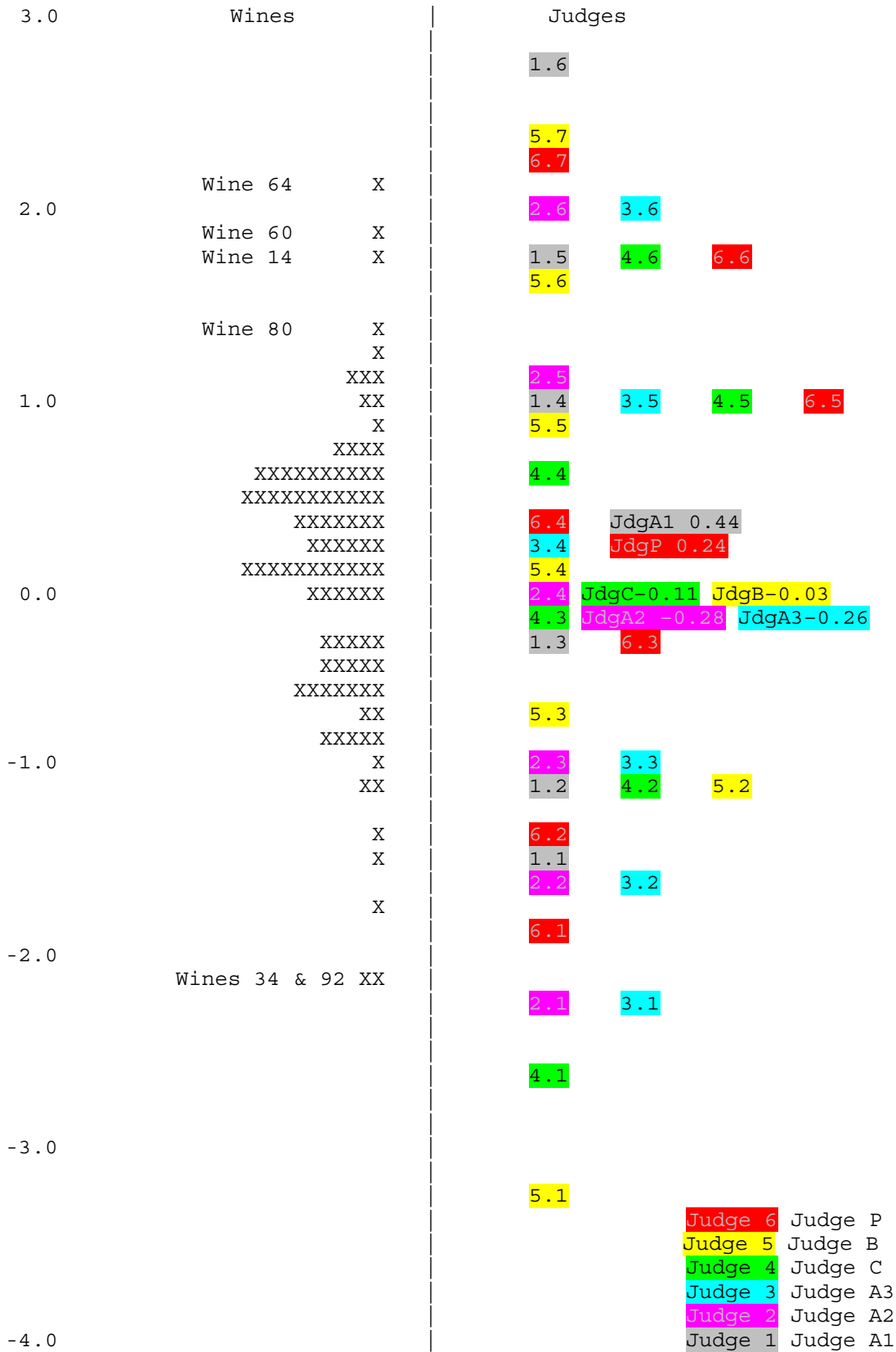
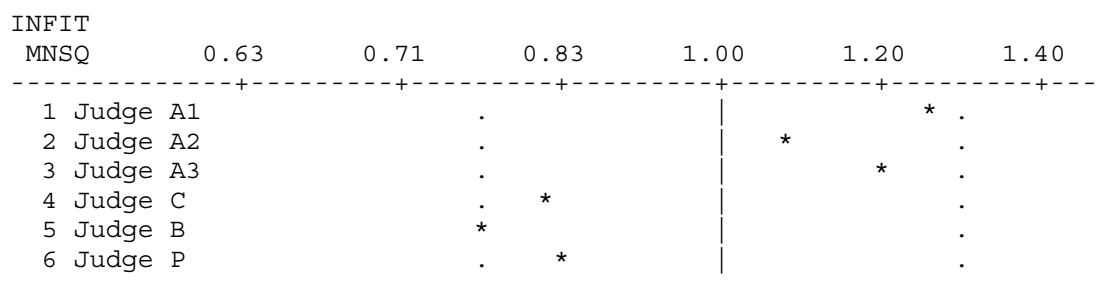


Figure 3. The performance of the wines and the thresholds of the judges, plotted on the same scale



**Figure 4. Infit mean square data for each judge**

**Table 2. The Infit Mean Square values and discriminations for each of the judges**

Item	Judge	Infit MNSQ	Discrimination
Item 1	Judge A1	1.25	0.50
Item 2	Judge A2	1.08	0.59
Item 3	Judge A3	1.08	0.56
Item 4	Judge C	0.82	0.73
Item 5	Judge B	0.76	0.77
Item 6	Judge P	0.84	0.73

This is further enhanced with the same data being analysed using only the expert judges' ratings. In this case, only the expert judges' scores were analysed. The results are displayed in Figure 5 and the results of the infit mean square for this analysis are shown in Figure 6. The infit mean square values are also shown in Table 3. What is clear from Figure 5 is that the various thresholds from each of the judges are well spaced out. What is especially noticeable is the measure of consistency given by the infit mean square display in Figure 6, which shows all of the experienced judges with results close to the ideal value of 1, reflecting excellent consistency. The infit mean square values and the discriminations for the expert judges alone are shown in Table 3.

**Table 3. The Infit Mean Square values and discriminations for the expert judges**

Item	Judge	Infit MNSQ	Discrimination
Item 1	Judge C	0.96	0.82
Item 2	Judge B	0.97	0.81
Item 3	Judge P	0.99	0.81

The consistency of the performances of these three judges is abundantly clear and well justifies the use of their results only in the final totals when the variability on fit and consistency for the other three associate judges is taken into consideration.

### The results for the wines – The case estimates

The results for each of the judges are brought together to estimate the performance of each of the wines. These are the results of the combination of the measures given by each of the three expert judges that take into account their consistent differences in harshness.

Table 4 shows the top 10 wines in order of their scores, both using the expert total score and the calculated score using the scaling system. The totals have been converted to a whole number by multiplying by two. The expert raw score is the sum of the converted scores of the expert judges on the scale of 0-7 as described above. The maximum score is the maximum total score awarded by the judges in the sample. It is noted that the order is unchanged using the expert rankings. In the all judge raw scores, the maximum awarded score was 38 and not 42. The inclusion of the associate judges in the Rasch scaling process has led, not surprisingly, to some slight alteration of

the order of the top 10 wines. Wines 80 and 60 received different total scores from all judges, but the process of transforming these scores into the 0-7 scale has given them the same value.

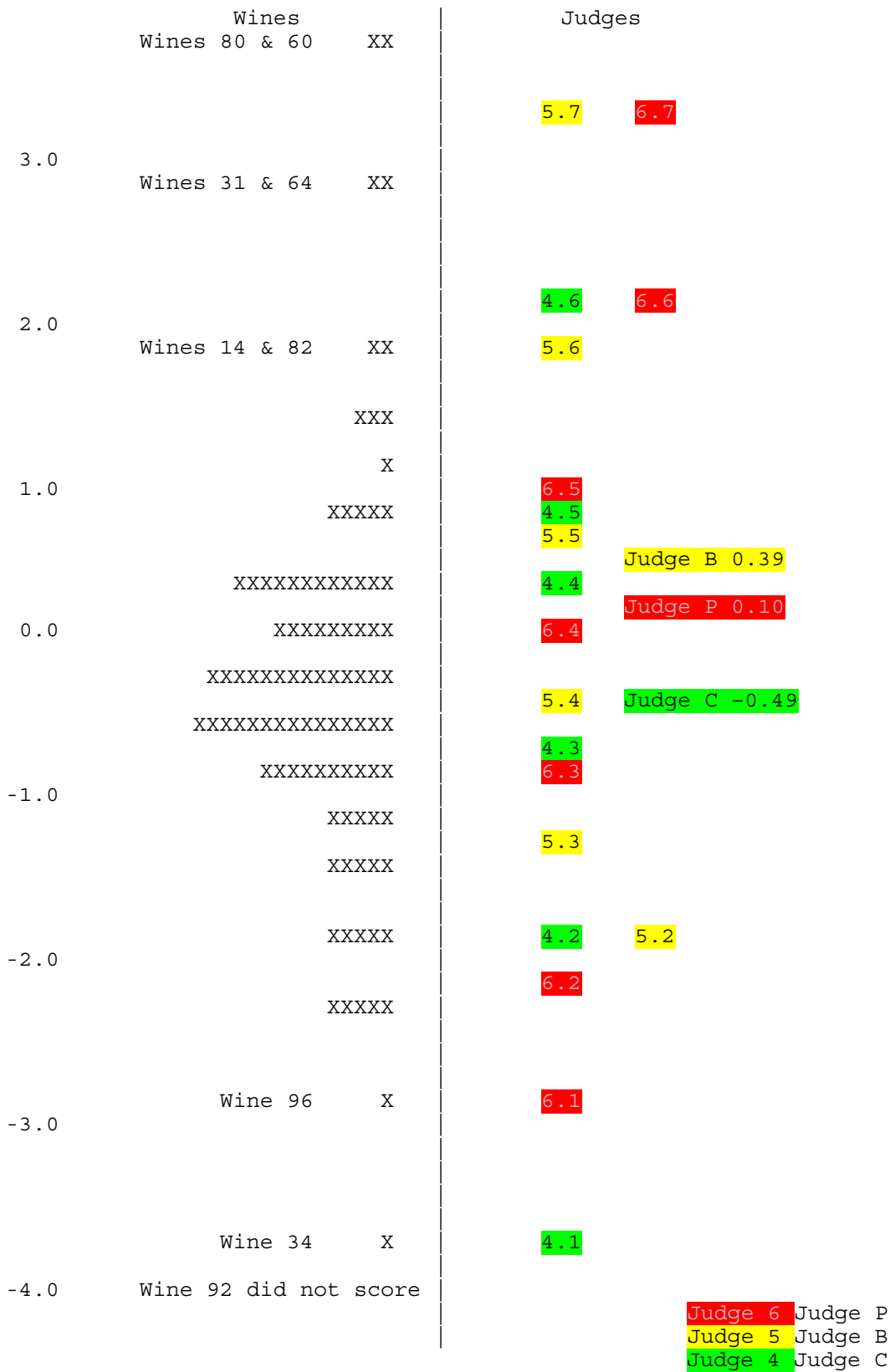


Figure 5. The wines and scores for the experienced judges only

INFIT		0.63	0.71	0.83	1.00	1.20	1.40
MNSQ							
1	Judge C				*		
2	Judge B				*		
3	Judge P				*		

**Figure 6. The infit mean square values for the expert judges alone**

**Table 4. The Ratings of the top 10 wines using the various rating systems**

Wine ID Code	All judge total 240	Expert total 120	All judge raw score	All judge maximum score on a scale 0-7	Expert raw score	Expert maximum score on a scale 0-7	All judge estimate using Rasch scaling	Expert estimate using Rasch scaling
80	205	112	29	38	18	19	1.37	3.70
60	211	111	32	38	18	19	1.86	3.70
31	204	111	27	38	17	19	1.10	2.87
64	215	111	33	38	17	19	2.05	2.87
14	210	106	31	38	15	19	1.68	1.83
82	199	105	27	38	15	19	1.10	1.83
41	201	105	26	38	14	19	0.96	1.42
61	197	105	24	38	14	19	0.70	1.42
A6	197	102	26	38	14	19	0.96	1.42
94	203	102	27	38	13	19	1.10	1.06

### The fit of the wines

The two ways in which Quest reports on the fit of items and persons, or in this case judges and wines are the infit mean square and the outfit mean square statistics. The infit mean square statistic tends to be used to investigate the fit of the items, in this case the judges, because it is calculated giving more weight to those wines closer on the scale to the judge than those further away. This is because those wines closer on the continuum to the judge provide more sensitive information. On the other hand, the outfit mean square statistic is not weighted and so is more affected by outlying scores. Just as the infit mean square statistic indicates when an item, in this case the judge, is being inconsistent, it can also happen that an individual wine may be judged inconsistently by the group of judges, where for some reason, there may be some disagreement about the individual wine. The outfit statistics indicate when there is a significant disagreement between judges and so is a reflection of outlying scores. In analysis of the scores of all judges, only five wines seemed to have some level of disagreement associated with them. The fit statistics for these five wines are shown in Table 5. In this table, the outlying scores are shown in bold.

There seem to be two reasons why these few wines are tending not to fit. In each case there is one score that seems very different from the others. In most cases, it is one of the inexperienced judges who seems to be at odds with the others. It is worth noting, however, that in two cases, wine 48 and wine 68, it is one of the experts who disagrees. In each case there has been an identified fault with the wine not picked by one or more of the others. Particularly interesting in this regard is the identification of the presence of volatile acidity (VA) in wine 48. The threshold of sensitivity to VA is said to vary greatly for different individuals. Table 6 shows the areas of disagreement for the expert judges, with the outliers shown in bold. There were only three wines out of the total of 98 and it can be seen that in each case there seems to be some wine fault at the heart of the problem.

**Table 5. The wines that do not fit into the Rasch model using all judges' scores**

Wine ID	Estimate	INFIT MNSQ	OUTFIT MNSQ	INFIT t	OUTFIT t	Judges' scores, associates row 1 experts row 2	Judge P's comments & score
19	-0.39	3.7	3.86	2.89	2.54	12.5, 16.5, <b>18</b> , 13, 13, 16	some cedar and earth, quite tannic full flavour (16)
31	1.1	2.97	3.23	2.66	2.39	<b>12.5</b> , 17.5, 16.5, 18.5, 18.5, 18.5	elegant fully ripened, well-composed, lingering tannins (18.5)
68	0.03	3.47	3.49	2.63	2.29	16, 13, 15.5, <b>18.5</b> , 14, 15	bitter astringent callow (15)
72	-0.25	3.2	3.37	2.47	2.23	12, <b>18</b> , 16, 15, 13, 15	a bitter taint, 2nd bottle, some ripe fruit under (15)
48	0.43	2.81	2.58	2.28	1.77	16, 17, 17, 15, 17, <b>12</b>	VA aldehyde (volatile acidity) (12)

This procedure could also be used to identify those wines that may be positively or negatively affected by a judge of questionable integrity, although there is no evidence of any lack of integrity in any of the judges in this analysis.

**Table 6. The wines that do not fit into the Rasch model using expert judges' scores**

Wine ID	Estimate	INFIT MNSQ	OUTFIT MNSQ	INFIT t	OUTFIT t	Expert judges' scores C, B, P	Judge P's comments & score
A9	0.14	5.58	6.19	3.22	2.91	14, <b>19</b> , 14.5	filled with oak and under-fruited (14.5)
68	0.14	4.81	4.78	2.9	2.44	<b>18</b> , 14, 15	bitter, astringent, callow (15)
48	-0.76	5.05	4.95	2.79	2.39	15, <b>17</b> , 12	VA aldehyde (volatile acidity) (12)

### A REDUCTION OF THE TOTAL NUMBER OF TASTINGS USING RASCH SCALING

The Cabernet Sauvignon class required 98 tastings, which must be a difficult and confusing task, even for the most experienced and conscientious judge. In the marking of essays, it has been possible to use the single scale established by the Rasch scaling process to connect the results of several markers and thereby take away the need for every marker to mark every script. Given the good fit of the Rasch model for the results of the wine judging, it follows that similar results could be obtained for the wines, in situations where there are experienced and well-trained judges, whose results fit the Rasch model. Accordingly, in this post hoc analysis of the data, each of the wines was assigned a random number and then they were ordered according to these random numbers. The wines were then divided into three groups, each of 32 wines, with extras, as shown in Table 7. The scores assigned by only two judges for each of the three groups were then analysed and the scores assigned by the third judge were ignored.

Thus, it was possible to assign randomly the wines as if they had been judged by only two instead of three judges, apart from the last two wines judged by all. It can readily be seen that instead of each judge judging 98 wines, they would only have been required to judge 66 wines. In this way it is possible to set up an assessment plan in which every judge rates at least some of the wines with each of the other judges and so it is possible to compare each of the judges with the others and

thus establish the harshness rating for each of them. An alternative process of dividing the wines up was also explored. In this second process, once again, each wine was assigned a random number and the scores for all judges for one quarter of the wines were included, two judges for the next quarter and so on as shown in Table 8. The important question that then arises is whether this reduced judging process can produce reliable results when compared to the complete process.

**Table 7. The allocation of the wines to judging teams, with wines divided into three groups, with two extra wines**

<b>Wine group</b>	<b>Judges</b>
Group 1 (wines 1-32 in list)	C, B
Group 2 (wines 33-64 in list)	B,P
Group 3 (wines 65-96 in list)	C,P
Group 4 (wines 97 & 98 in list)	C, B, P

**Table 8. The allocation of the wines to judging teams, with wines divided into four groups**

<b>Wine group</b>	<b>Judges</b>
1-26 (first quarter + 2)	Judges C, B & P
27-50	Judges C & B
51-74	Judges B & P
75-98	Judges C & P

**Table 9. The top 10 wines and their scores using various judge grouping**

Wine ID	All expert Score	All expert raw score	All expert estimate	Wine ID	3/4 expert estimate	Wine ID	2/3 expert estimate
80	112	18	3.7	80	Perfect	80	3.97
60	111	18	3.7	60	Perfect	60	3.26
31	111	17	2.87	31	Perfect	64	3.26
64	111	17	2.87	64	Perfect	31	3.15
14	106	15	1.83	14	2.47	14	2.72
82	105	15	1.83	82	2.44	82	2.45
41	105	14	1.42	41	2.47	61	2.24
61	105	14	1.42	84	2.47	71	1.75
A6	102	14	1.42	A6	2.35	41	1.75
94	102	13	1.06	68	1.76	49	1.75

Table 9 shows the estimates of each of the wines under various judging systems, using only the expert scores. The second column gives the experts' total score for each of the wines in rank order. The third column gives the experts' total raw score (on the eight point scale) and column 4 gives the performance estimates of each of the wines using the Rasch scaled score for the expert judges as discussed earlier. The remaining two sets of data show the estimates achieved using the reduced data. It can be seen readily that the groups of wines selected for the top 10 wines are consistent and it certainly seems to be a viable method for, at the very least, creating a short list for the final selection in any competition.

## SUMMARY OF THE QUEST ANALYSIS

Thus it has been shown that the Quest program can be used to place the performance estimates of the wines and the harshness levels of the judges on a single Rasch scale. Moreover, this scale can be used to examine the consistency of each of the judges and provide valuable feedback for each judge. It can also be seen that this may prove to be of value for judge training purposes. As well, it has been shown that the Rasch scaling procedure can be used to reduce the total numbers of wines tasted by each individual judge. At the very least this would enable a reliable selection of a short-list of finalists.

In summary, the scores from each judge out of a possible 20 points have been compressed into an eight point scale and these compressed score have been shown to fit the Rasch model well. This has allowed a ready comparison of the judges, both in terms of their harshness and their consistency. The analysis of these results provides excellent feedback for the training of the novice judge. Thus the Rasch scaling system provides a ready means for the analysis of such results.

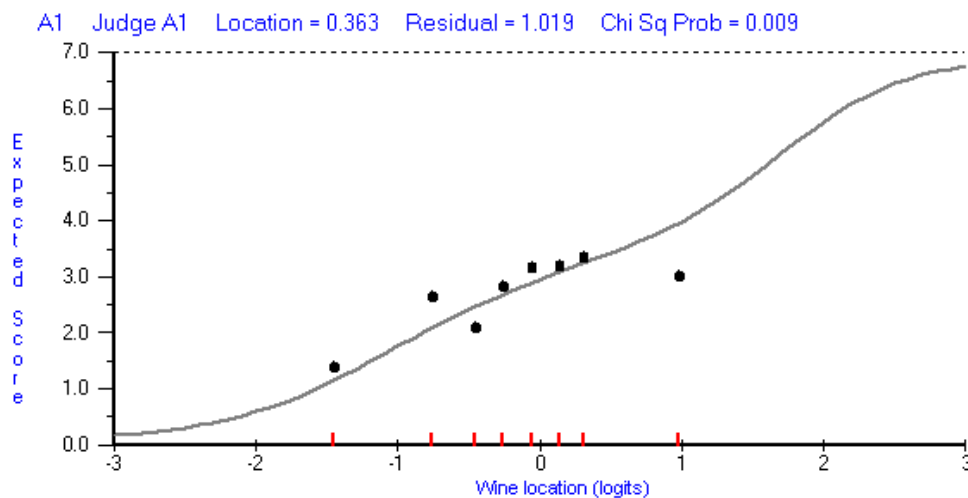
Moreover, it has been found that the Rasch scaling system, with its means of estimating the harshness of the judges and linking the scores of the judges with one another through common wines, has provided a procedure that can be used to reduce the number of wines needed to be tasted by each judge.

## THE RUMM PROGRAM ANALYSIS

### Analysis of the judges

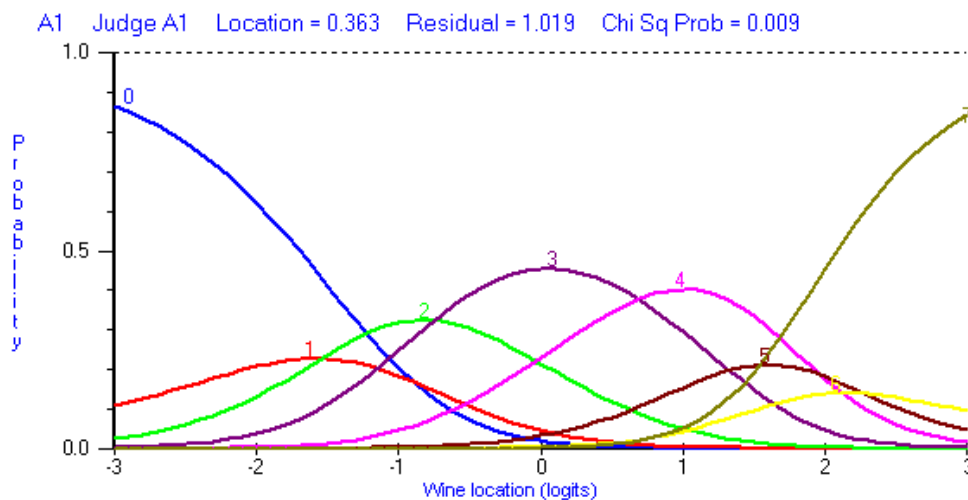
An alternative means of estimating the Rasch scale scores is the RUMM program (Andrich et al. 2000), which uses a pairwise conditional estimation procedure. The data were analysed using the RUMM program and it was found that there was good fit of the data to the model. This analysis allows for the provision of some valuable feedback of the judgment process. The item characteristic curve shows the eight point score against the location of the wines on the logit scale. The graph also shows group mean scores. The wines are divided into groups, in this case eight, and the means of the wine locations and the expected scores are calculated. These group mean points are represented as the dots on the graph of expected score against wine location. The idealised curve according to the Rasch model is plotted as well and a simple representation of how well the judges conform to the model is given by how closely the points lie to the curve. In addition, the residual and chi square statistics provided give an indication of the consistency of the particular judge. The RUMM analysis also provides category probability curves. In these curves the probability of a particular judge awarding each level from 0 – 7 is plotted against the Rasch scaled logit score. Clearly, in this graph, the ideal is to have each level being the most likely score awarded in order across the range, with each grade being the most dominant score for a section of the logit range. If the graph of a particular judge does not do this, it may reflect inconsistency or that some scores are under-used.

Figure 7 shows the item characteristic curve for Judge A1. The closeness of the plotted points to the curve reflects the closeness of the fit of the judge to the model. It can be readily seen that Judge A1, while conforming well to the model in the middle wine locations, is significantly off the curve in the lower and particularly the higher wine locations. In other words, Judge A1 is unreliable at the extremes of the scale. A low slope of the curve suggests low levels of discrimination. The large positive value of the residual and the small chi square probability score are indicative of the inconsistency of this judge. The low chi square probability reflects a large chi square statistic (14.81) which in turn reflects significant deviation from the model as evidenced in the wide spread of the points around the idealised item characteristic curve.



**Figure 7. Item characteristic curve for Judge A1**

These findings are well illustrated by the category probability curves for Judge A1 shown in Figure 8. It can readily be seen that a score of 1 or 2 is quite unlikely, with 0 or 3 dominating the logit range between  $-3$  and  $0.5$ . Similarly, a score of 5 or 6 seems very unlikely, with the scores of 4 and 7 dominating the upper logit range as shown in the item characteristic curve in Figure 7. These figures demonstrate that Judge A1 is not utilising the range as effectively as might be done and is not discriminating sufficiently well.



**Figure 8. Category Probability Curve for Judge A1**

The item characteristic curve shown in Figure 9 for Judge A2 shows more consistency than shown for Judge A1. The lower positive residual and the higher chi square probability are evidence of this. It can be noticed in Figure 9 that the plotted points lie much closer to the idealised curve, which is reflected in the high chi square probability. This tendency towards more consistency was reflected in the Quest program as well. The category probability curves in Figure 10 show a more consistent awarding of scores than Judge A1, with only a score of one being unlikely compared with the other scores.

Figure 11 shows the Item Characteristic Curve for Judge A3. As in the analysis using the Quest program, this analysis reflects moderate inconsistency from this judge with the relatively high

positive residual. However, like Judge A2, the plotted points lie close to the idealised curve leading to the higher chi square probability.

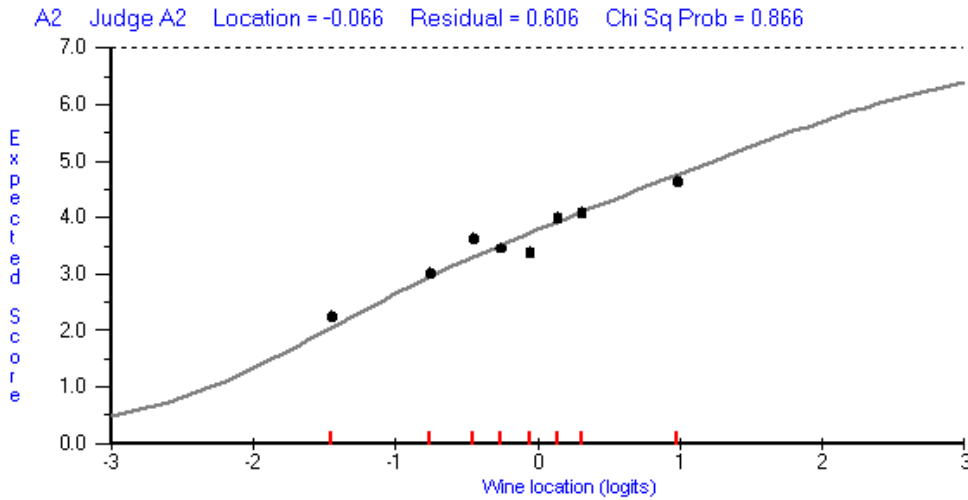


Figure 9. Item characteristic curve for Judge A2

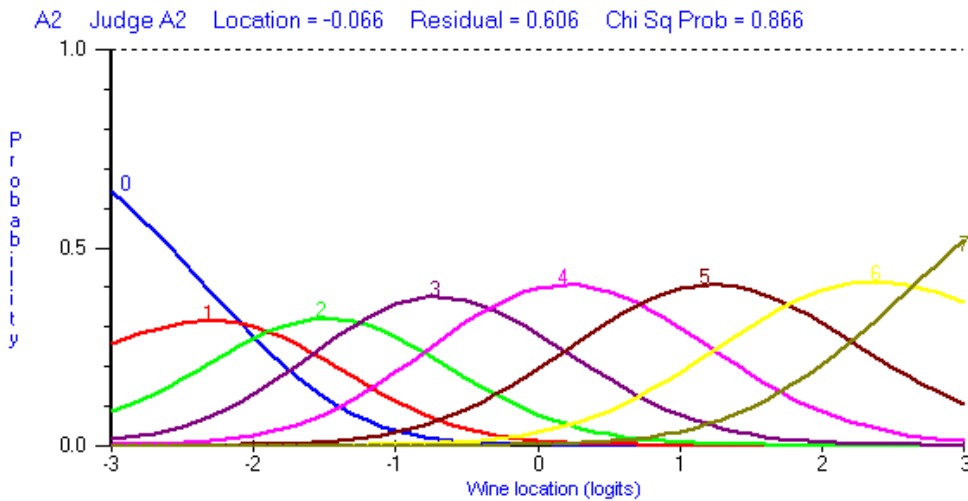


Figure 10. Category Probability Curve for Judge A2

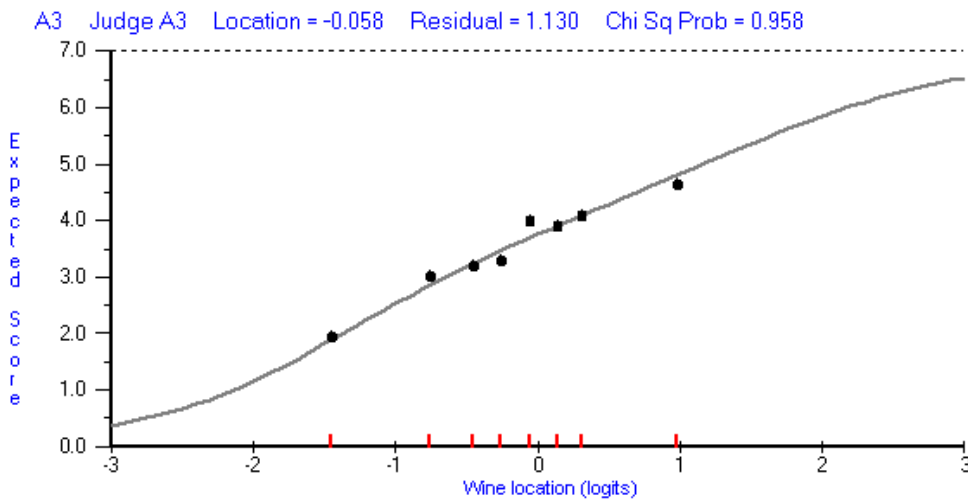
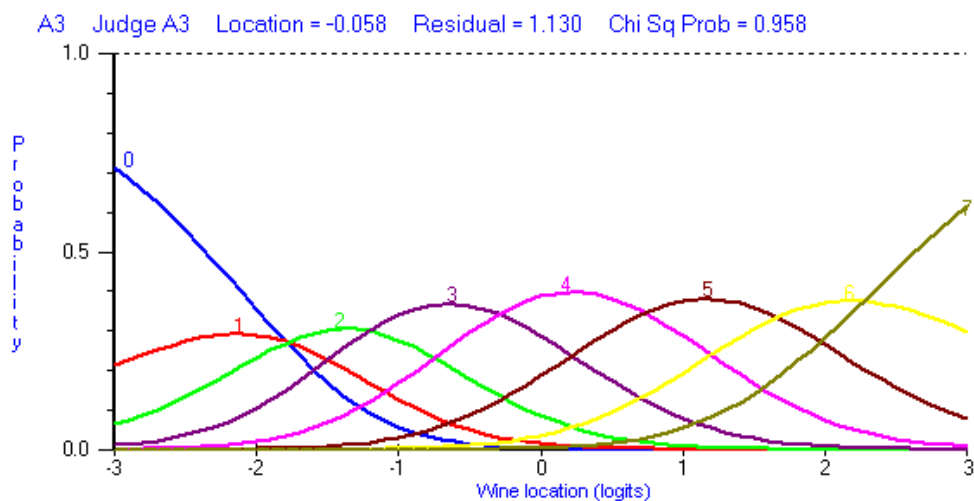


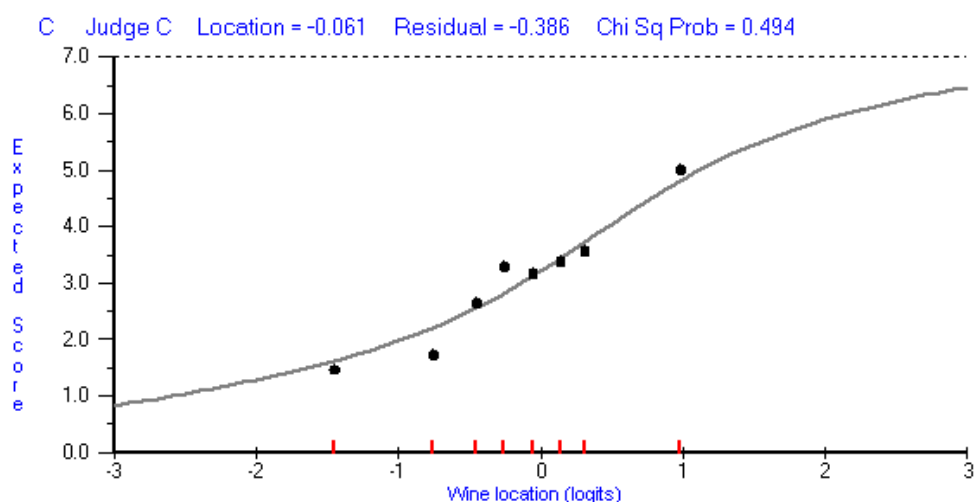
Figure 11. Item Characteristic Curve for Judge A3

The category probability curves shown in Figure 12 reflect a relatively even distribution of each score across the range, apart from a score of 1 which appears to have less likelihood across the range than scores of 0 or 2.



**Figure 12. Category Probability Curves for Judge A3**

Figure 13 shows the item characteristic curve for Judge C, one of the expert judges. The low value of the residual is immediately obvious and this is evidence of the greater level of consistency shown by this judge. This is also reflected in the category probability curves shown in Figure 14, which show a relatively even spread of scores across the range, although the score of 4 seems to be under-utilised, as does a score of 0. The slightly steeper slope reflects greater discrimination.



**Figure 13. Item Characteristic Curve for Judge C**

Figure 15 shows the item characteristic curve for Judge B, once again an expert judge. The consistency of this judge is shown in the low value of the residual, the closeness of the plotted points and the high value of the chi square probability. Similarly the steeper slope indicates good discrimination.

Figure 16 shows well the very even distribution of scores assigned by Judge B. Each of the scores is spread evenly across the range and this shows clearly the consistent awarding of scores.

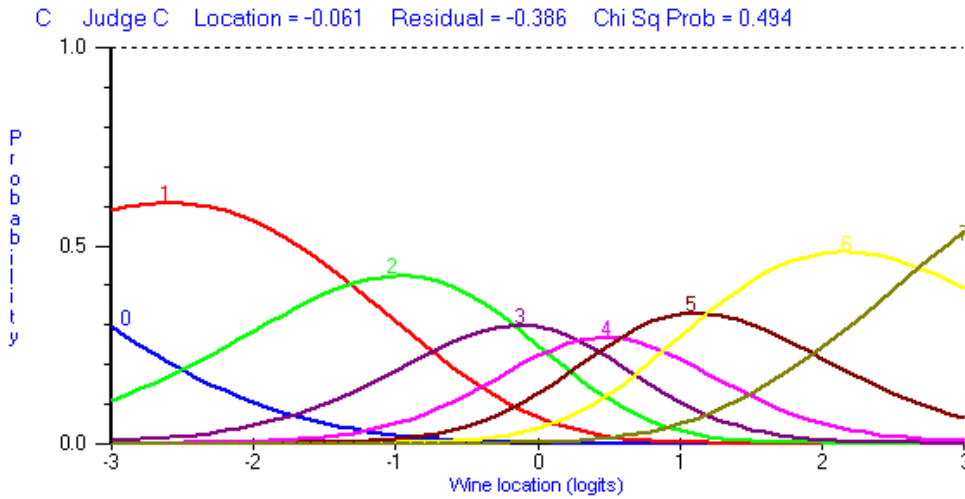


Figure 14. Category Probability Curves for Judge C

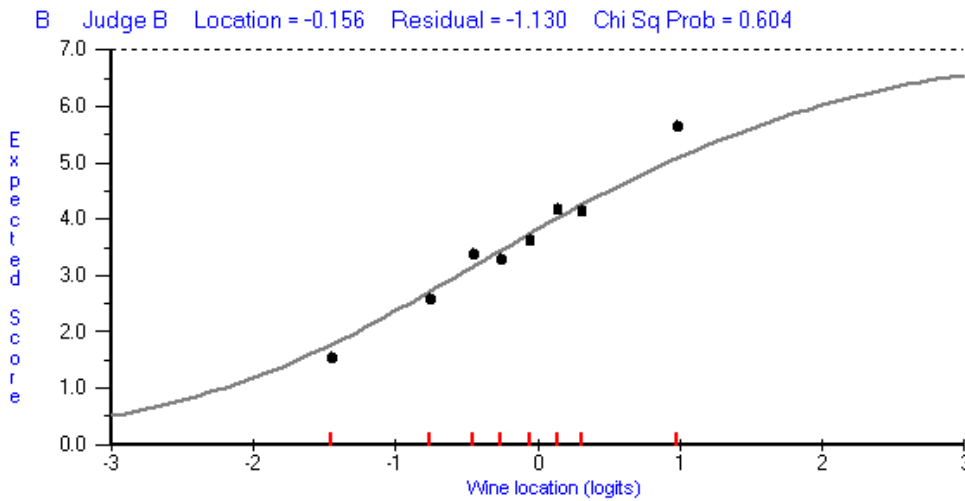


Figure 15. Item Characteristic Curve for Judge B

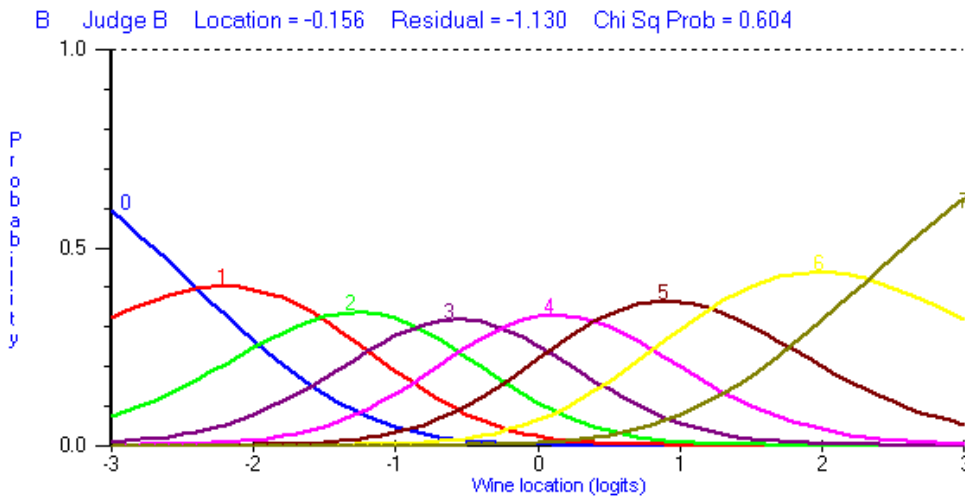
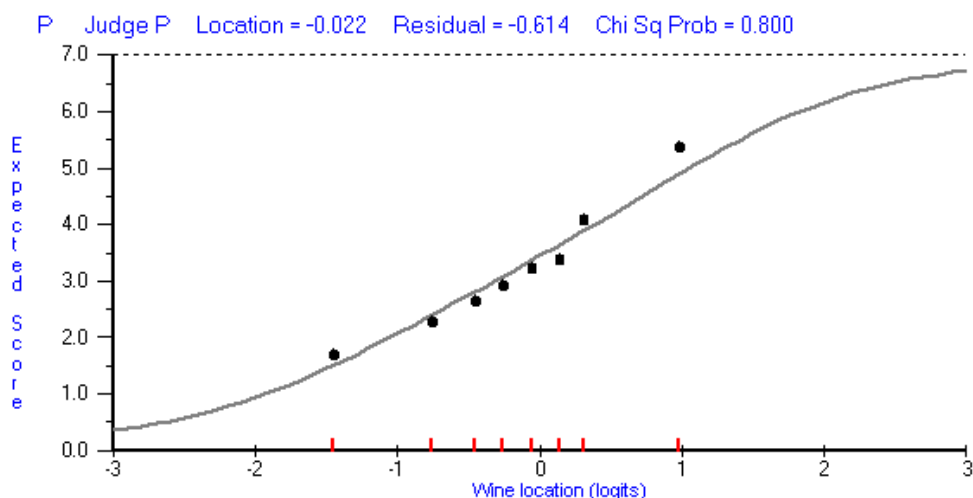


Figure 16. Category Probability Curves for Judge B

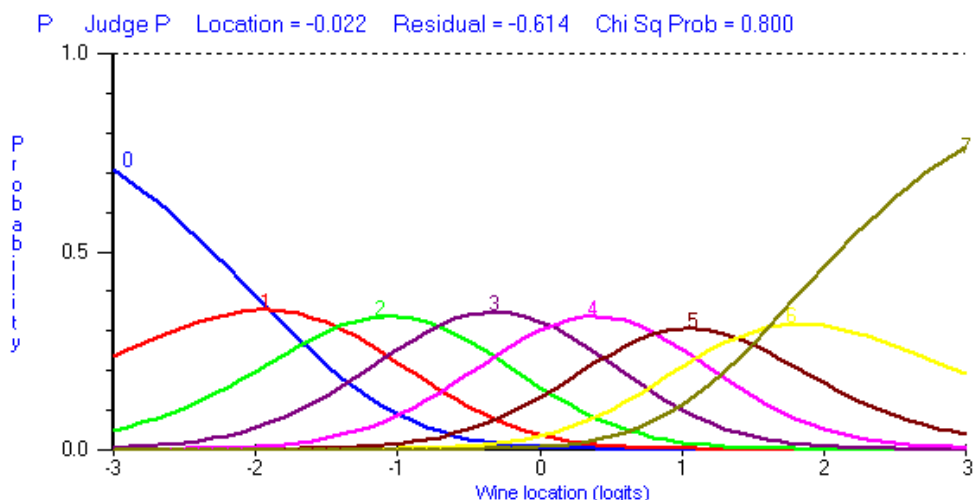
As has been shown for Judges C and B, Judge P shows consistency, both in the small spread of points around the idealised curve (and hence the high chi square probability) and in the low

residual score as shown in Figure 17. As for both the other expert judges, the steeper slope indicates good discrimination. The probability characteristic curves are shown in Figure 18.



**Figure 17. Item Characteristic Curve for Judge P**

Likewise, Judge P shows an even distribution of scores across the range, with the possible under-representation of a score of 6, with a score of 7 becoming more likely beyond the logit score of 1.6.



**Figure 18. Category Probability Scores for Judge P**

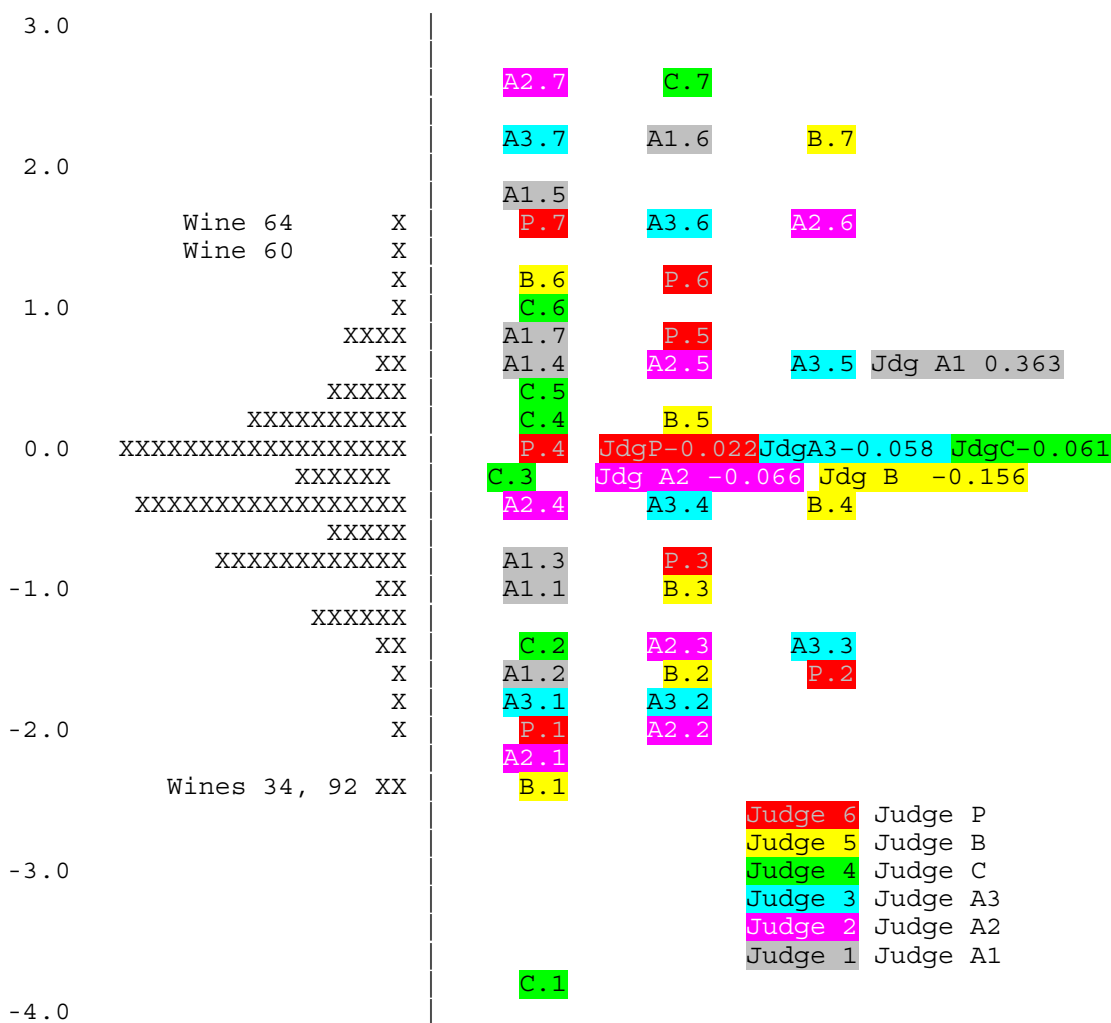
### Analysis of the Wines

Like the Quest program, it is possible to analyse the fit of the wines and see if there are any disagreements between the judges. The program identifies those judges for whom the residuals are above 2 or below  $-2$ . In the analysis using all judges, no wines were found to have residuals above 2, although as in the Quest analysis wines 48, 72, 68, 31 and 19 approached this mark. The program did identify that a few wines had a fit that was better than expected. These were six wines where there was very little variation between the judges. It is interesting to note the comments of Judge P for some of these wines: “clean and quite nice berry fruit”, “light style OK,” “oak far too charry”, “nice light going nowhere”, “rosé”, “savoury tart.” It is not surprising, then, that the judges found very strong agreement for these wines. A similar procedure was undertaken using the expert scores and interestingly, as well as those wines discussed above the

other group of wines on which the experts seemed to agree strongly were the very top wines that were awarded gold medals, underlining the effect of the considerable experience and training of these judges.

**The Item Map**

The RUMM program produces an Item and Person map as well, which in this case is interpreted as a judge and wine map. As in the Quest program, this shows the judges and the wines plotted on a single logit scale. The RUMM program produces as well an overall item location; in this case judge location and these have been added to the map shown in Figure 19. What is immediately clear from this is the unnecessary harshness of Judge A1 whose location is well above the other judges. It could be argued as well that Judge B is a little more lenient than the others. This is shown more clearly in Figure 20, in which the locations of each of the judges are plotted on a vertical scale. The outlying nature of Judge A1 is readily apparent.



**Figure 19. The Judge – Wine Map from the RUMM Program with the judges’ locations added on the right**

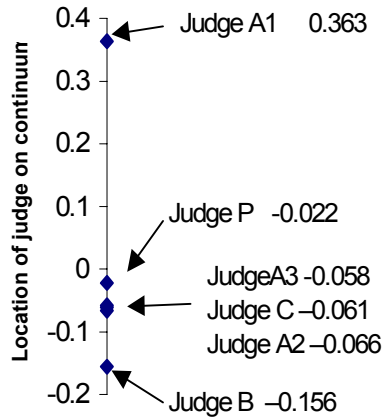


Figure 20. The relative harness of each judge plotted on a single scale

### A COMPARISON OF THE SCORING SYSTEMS

The procedures that have been used to rate the wines have been the two Rasch based procedures and the total scores. An important question to be asked is how do the various systems compare with one another? Correlations were calculated between the pairs of scores. There was a very strong correlation of 0.98 between expert total scores and the estimates using the Quest Program. Figure 21 shows a plot of the expert estimate using the Quest program against the expert total scores.

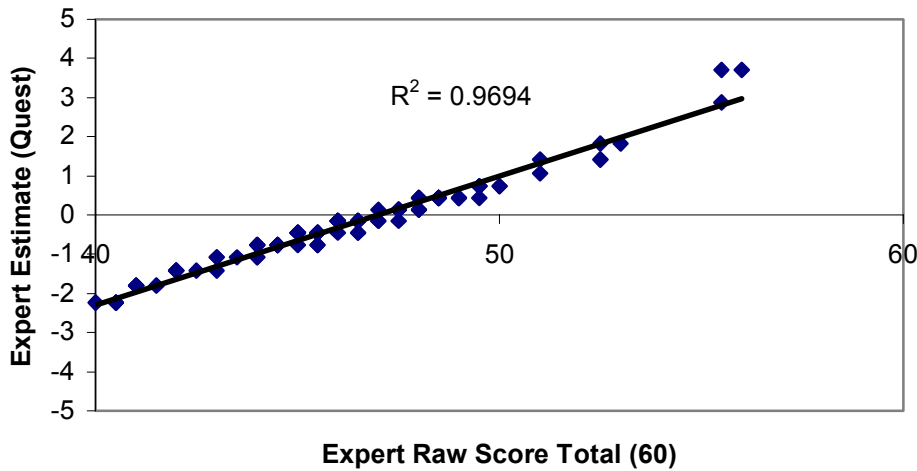


Figure 21. Expert estimate using the Quest Program against the expert total score

Likewise, there was a very strong correlation of 0.99 between the expert raw scores and the estimates using the RUMM program. Figure 22 shows the estimates from the RUMM Program plotted against the expert raw score.

The Quest and RUMM Program estimates yielded the almost perfect correlation of 0.999. The RUMM estimate is plotted against the Quest estimate in Figure 23.

Of particular significance here is the slope of the line. This slope of 0.919 indicates that the Quest scale and the RUMM scale are different. This difference is also borne out by the intraclass correlation coefficient of 0.89 that is noticeably less than the corresponding product moment correlation coefficient of 0.999. The Quest scale is more spread out than the RUMM scale. On the extremes of the scale, the Quest returns a value of 3.7 while RUMM gives 3.16. The difference in

scale seems to be due to the difference in estimation procedures used by the two programs and is somewhat puzzling. To summarise, the various estimates of the Rasch scaled scores correlate highly with the total scores and with one another. What is important, however, about the Rasch scaled scores is that they provide an interval scaled score. Such interval scores provide powerful means of making connections and comparisons.

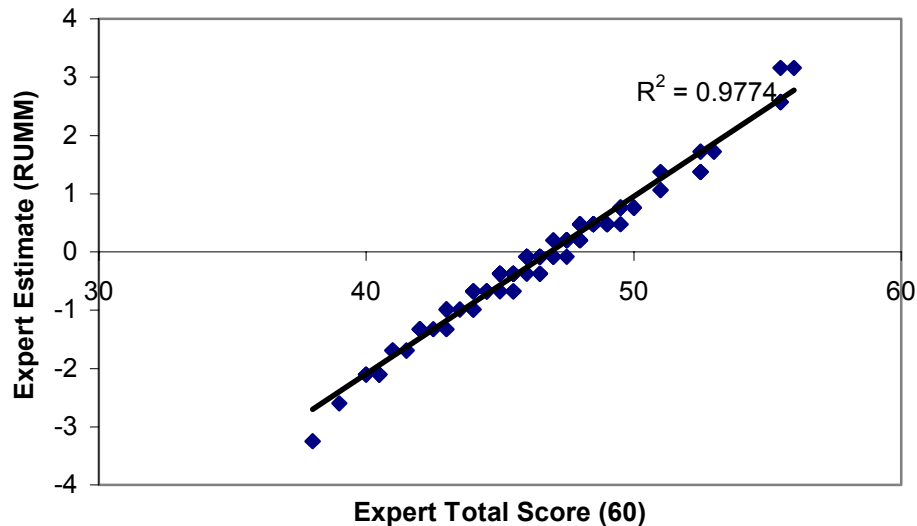


Figure 22. Expert estimate using the RUMM Program against the expert total score

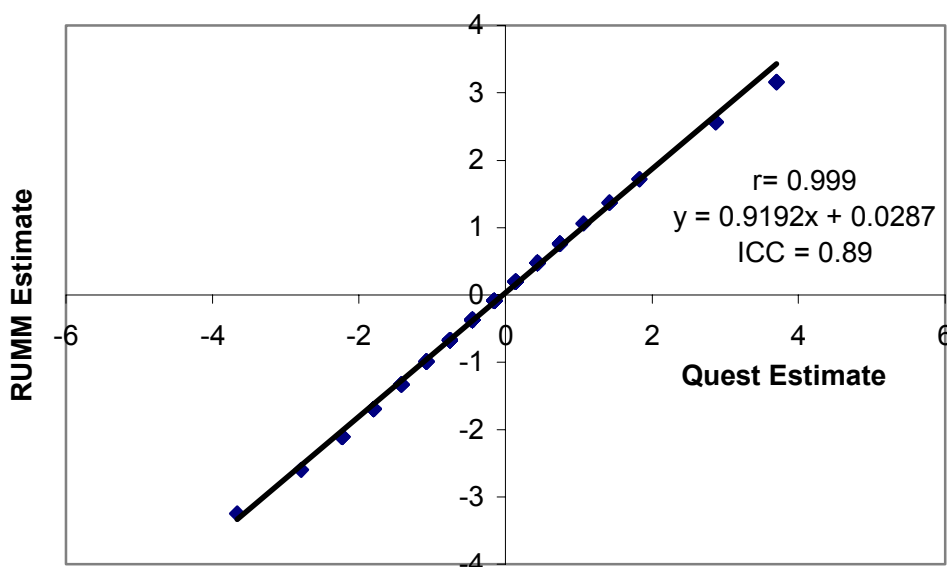


Figure 23. The RUMM expert estimate against the Quest expert estimate

Both the Quest and the RUMM programs provided estimates which fitted the Rasch model well and which gave good agreement with one another. For the purposes of feedback to the judging panel, however, the RUMM program, with its graphical displays provides easily interpreted feedback to those not familiar with the complexities of the Rasch scaling process.

In order to reduce the number of possible scores, these data were compressed from the 20 point scale, in increments of  $\frac{1}{2}$  point, to an 8 point scale from 0-7, compressing the scores in the lower range, making all scores below 12.5 zero and thereafter, compressing the range, with two possible scores sharing one final score. It has been shown that this compression makes little difference to the analysis. Thus, the Rasch scaling process could well be used as a means of simplifying the

difficult problem that arises when an attempt is made to apply too fine a scoring system, employing a wide range of possible scores, but actually using only a limited subset of them.

### CONCLUSION

The Rasch scaling procedures that have been explored have been shown to yield some useful insights into the judging of wine. The application of these educational measurement procedures represents an important use of these powerful measurement tools and indicates that in the wider community and business and sporting worlds there is much to be gained from the applying the results of educational research procedures. Of particular interest has been the ability to put the scores of the judges and the wines on a logit scale that has allowed connections to be made between judges, thereby providing a reliable means of reducing the total number of wines that need to be tasted by any one judge. Moreover, this method provides a means of determining both the harshness and the consistency of a particular judge. It has been suggested that this information would be invaluable for the training of judges giving easy to interpret feedback about how a judge is awarding scores.

### REFERENCES

- Adams, R. J., and Khoo, S. T. (1993). *Quest: The interactive test analysis system* [Computer software]. Camberwell, Victoria: Australian Council for Educational Research.
- Allen, M. (2000). Show business tasting politics are ripe for a coup. *Australian Magazine* November 25-26 2000.
- Andrich, D. (1999). Rating Scale Analysis. In G. N. Masters and J. P. Keeves (eds) *Advances in measurement in educational research and assessment*, (pp244-253), Oxford : Pergamon.
- Andrich, D., Lyne, A., Sheridan, B., and Luo, G. (2000) *RUMM2010* [Computer software]. Perth: Rumm Laboratory Pty Ltd.
- Bond, T. G., and Fox, C. M., (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah N.J.: Lawrence Erlbaum.
- Cronbach, L.J. (1964). *Essentials of Psychological Testing* London: Harper and Rowe.
- Linacre, J. M. (1999). Measurements of judgements in G. N. Masters and J. P. Keeves (eds) *Advances in Measurement in Educational Research and Assessment* (pp244-253), Oxford : Pergamon.
- Looney, M. A. (1997). Objective measurement in figure skating performance. *Journal of Outcome Measurement*, 1(2), 143-163.
- White, P. (2000). When it comes to awarding medals. *The Advertiser*, October 18, 2000.
- Wolf, R. M. (1997). Rating scales In J. P. Keeves (ed) *Educational Research, Methodology, and Measurement: an International Handbook* (2<sup>nd</sup> ed.), (pp. 958-965), Oxford: Pergamon.