COLOUR CODING SCALES AND COMPUTER GRAPHICS

A.M. Heath and R.B. Flavell

Department of Management Science
Imperial College of Science & Technology,
London, SW7 2BX, U.K.

ABSTRACT

Colour is a powerful tool when used to code information on a graphics display, and the growth in the number of colour displays increases the importance of understanding the potential of colour coding.

This paper describes the ways in which colour can be used to code information, and reviews previous guidelines. An experiment examining the quantitative, rather than qualitative, use of colour in ten different ordinal scales is described. The results generate new guidelines that demonstrate the importance of displaying a reference scale, and show that when applied out of context the previous guidelines can significantly underestimate the power of colour coding.

KEYWORDS: Human factors, Colour coding scales, Colour models.

INTRODUCTION

When displaying information graphically, colour is both extremely effective and efficient in terms of the speed and amount of data absorbed, and the level of recall. Colour coding does not, however, improve accuracy; if accuracy is important then numbers should be used. There are two principal methods of employing colour:

1. Nominal coding, in which colours are used qualitatively and are not ordered in any fashion but one colour represents one aspect of the information, e.g. a complex wiring diagram where each wire is represented by its own colour;

2. Ordinal coding, in which a (discrete or continuous) ordered scale of colour is used quantitatively to represent the ordered values of a set of data, e.g. a temperature scale where the coding is typically from blue (cold) to red (hot), or a contour map where heights above sea level are colour coded.

The research that has been carried out in the use of colour in visual displays has been into the application of nominal coding, especially in the area of enhancing visual search and identification performance under the auspices of US military contracts. In the 1950's the central question was "how many colours can be used for error-free recognition?" and the answers ranged from 6 to 10 depending upon the precise experimental conditions. This work was given impetus by the emergence of information theory as an analysis technique. In later years more complex experiments were constructed to examine the contribution of colour as merely one dimension out of several available (e.g. shape, size), towards visual search. For a survey of this work, see either Jones [5] or the later paper of Christ [1]. Most of this work however was concerned with the viewing of coloured physical objects, or in some studies with photographs, spectral lights or in one instance, with colour film.

In recent years, there has been a rapid development of colour raster graphics devices. The early ones were able to display 4 or possibly 8 colours, including black and white, and there was little control over the quality of the display. It is now possible to get raster graphics systems that select a pixel from a palette of over 16 million colours. The applicability of these early studies to colours displayed on a cathode ray tube (CRT) is dubious; the colour of physical objects is dependent on reflected light, whereas a CRT emits light and complicates the way in which colours are perceived.

Previously the cost of graphics equipment offering accurate colour representations restricted the use of colour to nominal codes, and the results generated for physical static displays remained satisfactory. However, with colour raster graphics devices, the cost argument is no longer valid, and ordinal scales flowing from one colour to another may be easily generated. The results for nominal scales do not hold when applied out of context to the use of ordinal scales, and as this technology is new it is not surprising that no research has been performed in this area.

The importance of such work grows when one considers dynamic displays; for example, moving a window around a map, displaying the on-line calculations of a model, or reporting on the current position of a number of devices. This paper describes experiments that have been performed to identify some guidelines for constructing ordinal scales. The results from the first of a series of experiments will be discussed. The direction of future work is also outlined.

The Definition of Colour

There are a variety of terms used to described the different aspects of colour. For consistency, the three terms 'lightness', 'hue' and 'saturation' will be used. For a fuller discussion of colour on CRT's see Murch [9] or Foley and van Dam [3, Chapter 17].

'Lightness'

If a light source illuminating an object contains all the wavelengths to which the human eye is sensitive, and the object reflects all those wavelengths equally, the colour of the object will either appear as black, or some intervening level of grey or white and is said to be achromatic. The lightness of the object, i.e. the amount of reflected light, determines the position in the achromatic scale.

If the strength of the light source is increased, the lightness remains constant, but its 'brightness' increases on a scale that runs from dark to bright. Colloquially the two phenomena are frequently treated synonymously and it may be difficult under some circumstances to distinguish between them. Because of the method of generating a colour display on a CRT, namely the emission of light and not reflection, these two phenomena are not independent.

'Hue'

Objects that reflect or emit unequal distributions of wavelengths are said to be chromatic. The hue depends on the dominant wavelength of the light and determines the name associated with the "colour" that we see, e.g. wavelengths between 450 and 480nm are predominantly blue, 500 to 550nm green and above 610nm red.

'Saturation'

Purity or saturation measures the spread of the distribution of wavelengths on a scale starting from the pure hue, i.e. fully saturated, then becoming less saturated until reaching a neutral grey. There are a number of systems for specifying the relationships between colours and they fall naturally into three classes.

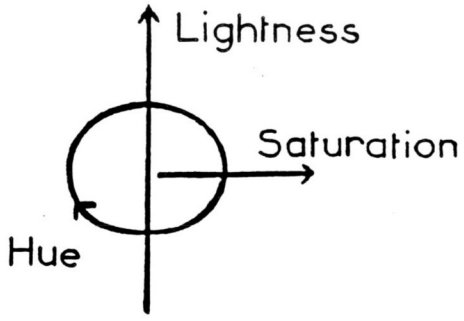
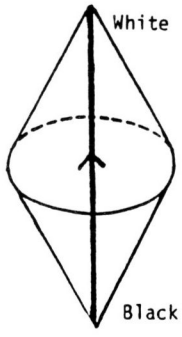


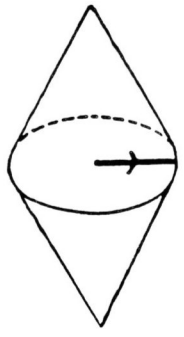
Figure 1a : The Perceptual Colour Dimensions.

i. Perceptual

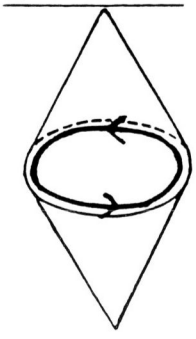
The earliest and probably best known is the Munsell system [10]. It consists of a set of standard colours organised in three-dimensions corresponding to hue, lightness and saturation (HLS); the colours are perceived as being equi-distant from their neighbours. The entire system is described in cylindrical co-ordinates, see Figure 1, in which the position of various scales is indicated. Figures 1b through 1d are only changing one dimension, 1e is changing both lightness and saturation simultaneously. A second perceptual system has been created by the Optical Society of America [11]. The use of these systems with graphics systems is considered by Meyer [7].



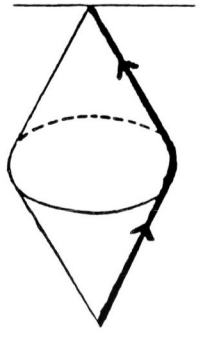
1b : Achromatic Scale



1c : Saturation Scale



1d : Hue/Rainbow Scale



1e : Coloured Scale

ii. Physical

The perceptual systems are purely subjective and in 1931 an objective system based on wavelengths was set up by the Commission Internationale l'Eclairage (CIE). This uses three primary imaginary colours to define all feasible colours.

iii. Graphics

Colour monitors create colours by combining primary colours. A CRT cannot produce all possible colours, and the limit or 'gamut' of colours it can produce is defined by the red-green-blue (RGB) colour model which is a cube with the three primaries on diagonally opposite corners, see Figure 2. Because the dimensions do not correspond to the perceived dimensions this model is difficult to manipulate and consequently user-orientated models, such as the HLS model, have been developed to map the RGB cube into more intuitive space. These mappings are one-to-one but not linear and can be done by a computer; see Foley and van Dam [3, pp611-620] for more details.

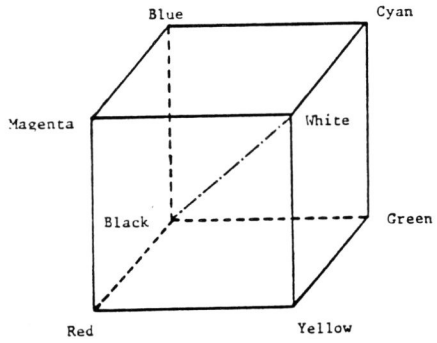


Figure 2: The RGB Cube

Experiments

It is an implicit assumption with the use of ordinal scales that people can look at individual colours from within a scale and be able to locate their relative positions in the scale. It is also implied that if the scale represents some physical property, e.g. height on a contour map, then people are capable of subjectively translating the various colours into appropriate values of the property.

It was decided to start by looking at these fundamental assumptions under very simple environmental conditions. A high resolution colour graphics device (PLUTO [13]) capable of generating more than 16 million colours was used, controlled by a low-level version of GKS implemented on a SAGE IV microcomputer. The experiments were restricted to static displays of colour placed on a mid-grey background.

The initial experiments were designed to investigate the errors people made when trying to place isolated colours into their correct place in an ordinal scale. Four different types of scale were constructed, corresponding to Figure 1b to e respectively;

i. achromatic or grey scale; starting with black,then passing through grey to white

ii. saturation; ranging from grey to a pure colour

iii. hue or rainbow; starting with white then running through the pure colours, in a rainbow sequence, ending with red.

iv. coloured; ranging from black through a pure colour to white

The rainbow scale has no intuitive beginning or end, and yet is probably most widely used in computer graphics; the sequence suggested by Poulton [14] was implemented. There were four different variations of scales ii and iv, based on pure colours of red, blue, green and yellow. Each scale consisted of twenty colours, the individual colours were generated by initial matching under controlled lighting conditions with the Munsell colours.

The advantage of using Munsell colours is that they ensure equal perceptual spacing although in practice the spacing was modified slightly to overcome induced effects from the background. The length of the scales was selected to be longer than that required for error-free performance (on the basis of the previous work and a few initial trials) and was held constant in this series of experiments.

The experimental procedure was as follows for each subject:

1. A colour blindness test [6] was performed

2. A scale was displayed on the graphics screen whilst instructions were given on a neighbouring green monochrome screen; apart from a shaded light for the keyboard, there was no other light in the windowless room. The subject controlled the rate of instructions with a numeric keypad. The scale was 20 uniform rectangles, each 12mm x 15mm in size and arranged horizontally at the top of the screen separated by a grey border. The colours were numbered underneath from 1 to 20. The computer displayed another rectangle measuring 16mm x 20mm in the centre of the screen. Five examples were given, in which this lower rectangle was filled with a colour from the scale and the correct number was shown underneath. Then another five practice colours were generated and the subject typed in the (subjectively) appropriate number and the computer then responded with the correct one. This was all the training that was received.

3. A sequence of 60 random colours from the scale were now displayed in the centre of the graphics screen. There was a gap of 77mm between the bottom of the reference scale and the top of each of these colours. The colours were shown alone or in groups of either 5 or 20. For each colour the subject had to type the appropriate number. No time limit was placed on the subject to respond, although unbeknown the response time was recorded by the computer, and there were facilities for the correction of errors. The computer did not respond with the correct number.

4. The horizontal reference scale was removed and the subject repeated stage 3 with a different random sequence of 60 colours.

5. Stages 2 through 4 were then repeated for each of the four types of scales. Each subject only did a single variation of the saturation and coloured scales with an equal number of people doing each variation.

Thus each subject made a total of 480 responses. The total response times ranged from 30 minutes to an extreme of 140 minutes. No subjects were retested, and motivation was generated by offering prizes for lowest error performances.

Results

The experiments were designed to see if people could respond correctly to a colour and place it accurately in an ordinal scale. The criterion is therefore the degree of error involved in this response. The errors made by subjects can be considered to be of three types. Firstly, procedural errors caused by bad typing or confusing the ends of the scale and responding 1 to 20, or 20 to 1. Secondly, perceptual errors due to induced effects of neighbouring colours either in the scale or in the main display area, perceptual errors may also be caused by the difference in size between the stimulus and the colours in the reference scale. The final group of errors are judgemental i.e. comparison errors when the reference scale was shown, combined with errors of memory when the scale was removed. The following analysis includes all types of errors; there were only 13 procedural errors in over 16,000 responses.

A simple measure would be to take the percentage of correct responses but this conceals a lot of information such as the size of error, and the uniformity of error on the scale. The responses were converted into percentage error terms - on a 20 colour scale, an error of 1 is $100/(20-1) = 5.3\%$ - and the results for a total of 33 adult subjects shown in Appendix B and Figure 3.

The variations of the coloured and the saturated scales are all clumped together; there was no significant difference between the coloured scales, however, the saturation scales were significantly different at the 5% level using a rank-sum test. The yellow saturation scale was the best, followed by the green, blue and then worst of all the red saturation scale. It is unclear whether this result holds for all saturation scales, or if it is simply caused by the specific combination of colours used in this experiment.

As is clear from Figure 3, the penalty for not displaying a reference scale is extremely high, on average doubling the size of error. This is most marked in the rainbow scale, which lacks an intuitive internal structure, making it difficult to associate magnitude with the scale. Displaying different numbers of colours clearly showed that they were being used for internal reference purposes; the performance with 20 was better than with 5 which in turn was slightly better than with 1.

No significant difference was found between males and females but there is a significant difference between those with normal colour vision and those who made mistakes in the colour blindness test. This is despite the fact that four or less mistakes is not considered as evidence of colour blindness. One might consider the mistakes and the poor performance both result from sheer carelessness but if this were the case, these people should have completed the tests more rapidly, which they did not. This suggests a physical rather than psychological reason.

A red-green colour blind person has subsequently taken the tests (but too late to be included in the analysis) and scored in the top third! This may indicate the subject made a conscious effort to compensate, or that the difficulty of the task dominated any errors due to colour matching, the reason is unknown as the sample size is obviously too small.

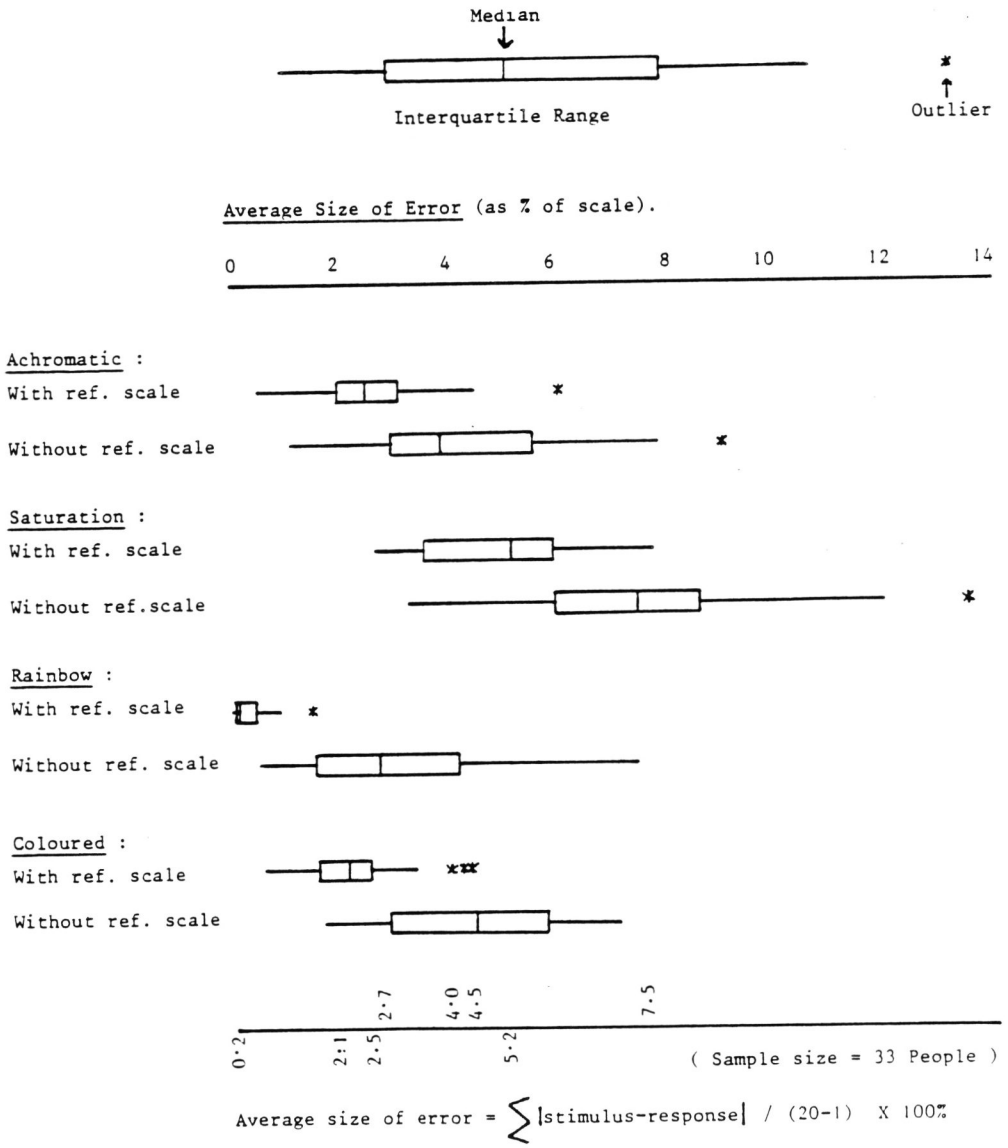


Figure 3: Box Plots to show distribution of Average Size of Error.

Although two subjects spent considerable time memorising the colours and performed very well, there was no significant correlation between the total response time and the error rate. When the responses within each scale are examined, it is found that the variance of errors increases towards the middle of the scale. This was despite the fact that each colour was selected as being equidistant (as defined by Munsell) from its neighbours. The result is not entirely surprising, especially for scales with natural beginnings and/or endings, which provide subjects with reference or anchoring points.

Finally, information theory was used to measure the real number of separate colours people were using in the scales. The method of analysis is described in Appendix A; the results are shown below in Table 2.

The use of the rainbow scale without a reference scale is equivalent to the early experiment on nominal scales, and as expected this result agrees with previous work, cf. Miller [8]. When a reference scale is shown the coloured scales are all significantly better than the achromatic scale at the 5% level using a rank-sum test. However, after the reference scale was removed, the achromatic and coloured scales performed equally well. This is interesting as of course the achromatic scale may be produced on a black-and-white system.

Scale	Number of colours used S_t (rounded)	
	With ref.scale	Without ref.scale
Achromatic	8	6
Saturation		
-Red	5	3
-Blue	5	4
-Green	5	4
-Yellow	7	4
Rainbow	17	7
Coloured		
-Red	11	6
-Blue	9	5
-Green	9	6
-Yellow	10	6

Table 2: Number of individual colours used in each scale

Note: There was no significant difference between the coloured variations.

Discussion and Future Work

The experiments described here are only preliminary and other experiments are currently continuing. The length of the scale was held constant at 20 for all of these experiments; it is hypothesised that if the length were reduced a maximum number of usable colours would be found and this is being investigated.

Whilst the colours were equidistant in perceptual space the errors were not uniform. It is desirable in practice to have for each stimulus an approximately equal variance, and new scales of equal discriminability are being constructed based on the methodology of Garner and Hake [4].

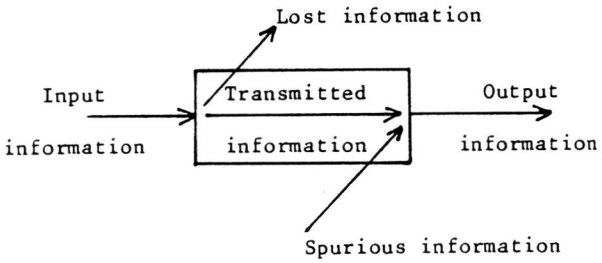
Another hypothesis under investigation is that scales of very uneven discriminability might increase the number of usable colours and that in fact, there is a trade-off between uniformity of error and maximum transmission of information. No time limits were placed on the response time; it is hypothesised that the rainbow scale would be less useful if there was a limit. This has obvious implications for dynamic displays.

These results for ordinal scales differ from the previous experimental results for nominal codes. This highlights the need to ensure experimental results are not applied beyond the context in which the experiment is performed. Of course, this statement also holds for the experimental results presented here. Consider for example, a colour coded contour map. Phillips [12] has shown that the performance of scales depends on the task being performed; a rainbow scale is best when a spot height is needed, whereas a coloured scale is better when the direction of the slope is of interest.

The objective of this work is to establish guidelines for the use of ordinal colour scales for display purposes, especially computer graphics. The first results measure the performance of different colour scales both with and without a reference scale, and further hypotheses are suggested. Much more work is needed to examine the quantitative rather than qualitative use of colour coding, especially when used in dynamic situations.

Appendix A: Information Theory

This appendix summarises the relevant information theory that has been used in the main text; for more details see Sheriden and Ferrell [15, Chapter 5 and 6], or Edwards [2]. The principle is shown in the diagram below, namely that there may be differences between information that is sent and information received, that these differences are due to information being lost or noise being added, and that therefore the useful message is what is transmitted. Psychologists have been applying this form of analysis since the 1950's to situations involving stimuli and responses.



A measure of information is called the 'bit'; this measures in effect the number of binary discriminations that must be made in order to specify one event from a number of alternatives. For example, if one is presented with a colour drawn uniformly randomly from a scale of eight, i.e. with a probability p=0.125, 3 discriminations are needed to identify the colour, the first to select which of the two groups of four colours, the second which of the two groups of two and the last which of the two remaining colours. Mathematically, the amount of information needed (I) is:

$$I = -\log_2 p = 3 \text{ bits}$$

If the selection of a colour were not uniform but according to a discrete distribution with probabilities p_i for the ith colour, a similar argument yields:

$$I = -\sum_i p_i \log_2 p_i$$

Consider now an experiment in which there is a scale of S colours each of which is presented $N_{.k}$, k=1..S, times randomly. A subject responds to each presentation by stating which colour he perceives it to be; for the $N_{.k}$ presentations of colour k, let the responses be N_{jk} perceptions of colour j, j=1,...S. Note that some of these cells may be zero. Let $N_{j.}$ be the total number of perceptions of colour j and obviously:

$$\sum_k N_{.k} = \sum_j N_{j.} = \sum_{jk} N_{jk} = N$$

If there are a large number of presentations, the ratios $N_{.k}/N$, etc may be interpreted naturally on frequency probabilities (the large sample size is crucial when using psychological data). Given the discussion above, the information in the stimuli is:

$$I_s = -\sum_k p(k) \log_2 p(k) \text{ where } p(k) = N_{.k}/N$$

and the information shown in the responses (which may include noise) is:

$$I_r = -\sum_j p(j) \log_2 p(j) \text{ where } p(j) = N_{j.}/N$$

The purpose of such an experiment is to determine what information is conveyed by a stimulus. Assume a stimulus k engenders response j, by a similar fundamental argument to above:

$$I(k|j) = \log_2 p(k|j)/p(k)$$

i.e. the information conveyed is dependent upon the prior probability of k before j were observed and the changed posterior probability for k after j occurs. From a series of responses the average information transmitted is:

$$T = \sum_{jk} p(k,j)\ I\ (k|j) \text{ where } p(k,\ j) = N_{jk}/N$$

assuming each pair (k,j) is independent. By the use of Bayes theorem, this may be rewritten easily as:

$$T = I_s + I_r - I_{rs}$$

where $I_{rs} = p(k,j) \log_2 p(k,j)$

The lost and spurious information are defined naturally by balances as:

$I_s - T$ and $I_r - T$ respectively.

If there was a perfect correlation between stimulus and response, then $T = I_r = I_s$. This relationship may be used to determine how many different stimuli were actually being used accurately, i.e.

$$S_t = 2^T \text{ (generally rounded to integer).}$$

Appendix B - Results

OBSERVER NUMBER	PERCENTAGE OF CORRECT RESPONSES	AVERAGE SIZE OF ERROR	SEX	SCORE IN COLOUR BLINDNESS TEST	TOTAL RESPONSE TIME (MINUTES)
19	40.83	6.05	M	4	33.45
31	43.75	5.37	F	1	46.65
4	45.62	5.16	M	2	38.02
27	46.85	5.26	M	0	74.43
5	48.12	4.53	M	1	83.32
12	48.54	4.47	F	0	32.15
20	48.96	4.53	M	1	57.90
29	49.17	4.00	M	0	35.25
24	49.37	4.05	F	0	64.33
15	49.58	4.47	M	0	63.68
6	49.79	4.11	M	0	29.53
33	50.83	4.63	F	0	46.60
9	51.04	4.42	M	0	46.87
7	53.12	3.74	M	0	47.72
25	53.75	3.95	F	2	51.58
28	54.17	3.63	M	2	65.32
13	54.37	3.47	F	0	28.63
22	54.79	3.63	F	0	50.55
8	55.00	3.21	M	0	45.38
21	55.83	3.53	M	1	85.03
1	56.04	3.42	M	0	44.70
2	57.29	3.89	M	0	46.73
32	57.29	3.53	M	0	47.65
16	58.54	2.95	F	0	45.57
26	58.96	2.95	F	0	48.70
17	59.17	3.16	M	0	39.63
11	60.62	3.32	F	0	49.02
23	61.87	3.00	M	0	46.93
3	62.29	3.11	F	0	42.07
30	66.67	2.37	M	0	140.32
10	68.33	2.05	M	0	67.77
14	73.33	1.79	M	0	45.08
18	80.63	1.32	M	0	97.42

REFERENCES

1. Christ, R.E., Review and analysis of colour coding research for visual displays. Human Factors, 17(6), 542-570, 1975.

2. Edwards, E., Information Transmission, Chapman & Hall, London, 1964.

3. Foley, J.D., and Van Dam, A., Fundamentals of Interactive Computer Graphics, Addison-Wesley, 1982.

4. Garner, W.R., and Hake, H.W., The amount of information in absolute judgements. Psychological Review, 58, 446-459, 1951.

5. Jones, M.R., Color Coding. Human Factors 4(6), 355-365, 1962.

6. Ishihara's Tests for Colour-Blindness, Kanehara & Co.Ltd., Tokyo.

7. Meyer, G.W. and Greenberg, D.P., Perceptual colour spaces for computer graphics. Computer Graphics 14(3), 254-261, 1980.

8. Miller, G.A., The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 63(2), 81-97, 1956.

9. Murch, G.M., The effective use of color, TEKniques 7(4), 13-16, 8(1), 4-9, 8(2), 25-31, 1983/4.

10. Munsell, A.H., Munsell Book of Color, Baltimore, Munsell Colour Company Inc., 1967.

11. Nickerson, D., OSA uniform color scale samples: A unique set. Color Research and Application, 6(1), 7-33, 1981.

12. Phillips, R.J., An experimental investigation of layer tints for relief maps in school atlases. Ergonomics 25(12), 1143-1154, 1982.

13. Pluto Power! Personal Computer World, p126, December 1982.

14. Poulton, E.C., Colours for Sizes, Applied Ergonomics, 6(4), 231-235, 1975.

15. Sheriden, T.B., and Ferrell W.R., Man-Machine Systems, MIT Press, London, 1981.