Screen complexity and user design preferences in windows applications

Tim Comber
Southern Cross University

John R. Maltby
Southern Cross University

SCREEN COMPLEXITY AND USER DESIGN PREFERENCES IN WINDOWS APPLICATIONS

T. Comber and J. R. Maltby

Centre for Computing and Mathematics

Southern Cross University

Lismore

New South Wales 2480

Australia

E-mail: {tcomber, jmaltby}@scu.edu.au

ABSTRACT

This paper evaluates the validity of a formal method for assessing the quality of screen layouts in graphical user interfaces. A technique developed by Bonsieppe for quantifying the layout complexity of a printed page has been applied to the opening screens in thirteen Microsoft Windows applications. Thirty subjects were asked to rank the same thirteen screens on the basis of "good" design. A significant negative correlation was found between the subjects' rankings and the complexity ratings, indicating that users' do not like "simple" screens. The reasons for this negative correlation are explored.

KEYWORDS

GUI, Windows, screen, design, layout, complexity

INTRODUCTION

This paper attempts to evaluate a quantitative method for determining the relative order and disorder of objects in a graphical user interface (GUI) by comparing the output of the method with users' perceptions of good screen design. The method itself is based on the Shannon formula for entropy [1] and has been used for typography by Bonsieppe [2], for text screens by Tullis [3][4] and for Microsoft Windows application screens by Comber [5] and Maltby [6].

Theory

Bonsieppe identifies two types of order in typographical design: system order and distribution order. System order is determined by classifying objects according to common widths and common heights and distribution order is determined by classifying objects by their distance from the top of the page and from the side of the page. The proportion of objects in each class is then used to determine the complexity C of the layout from the formula

$$C = -N\sum_{i=1}^{i=n} p_i \log_2 p_i \qquad \text{. . . (1)}$$

where:

$$C = -N\sum_{i=1}^{i=n} p_i \log_2 p_i$$

and

$$p_i = \frac{n_i}{n}$$

where:

N = total number of objects (widths or heights, distance from top or side of page)
n = number of classes (number of unique widths, heights or distances)
n_i = number of objects in the *i*th class
p_i = proportion of the *i*th class.

Bonsieppe tested the formula by comparing two versions of a printed catalogue: the original design and the design modified according to Bonsieppe's theories on layout. The new version was found to be 39% more ordered than the original. This figure was considered to be supported by the fact that the new version *appeared* to be more ordered than the original.

Screen design guidelines frequently recommend that design goals should be to minimise the complexity of a display or make screens as predictable as possible [7][8][9]. In line with these recommendations, Tullis [3] has proposed that a modified version of Bonsieppe's technique, which excludes system order and identifies objects as data and labels, is an objective method of assessing screen design. Tullis applied this technique to screens that had been identified as either narrative or structured. The structured screen returned a lower complexity figure than the narrative screen, in agreement with the finding that users performed better with the former.

Previous work [5][6] has indicated that the Bonsieppe technique does provide a meaningful measure of layout complexity in GUIs. The work done by Tullis considered distribution order only. Tullis felt that system order was not applicable to display screens which consisted of text alone. Since GUI applications display both text and graphics, system and distribution order would both appear to be appropriate metrics. However, Tullis did note a problem when analysing text screens which was found to be accentuated when examining Microsoft Windows applications. The problem relates to the difficulty in determining the extent of an object. A typical Windows screen contains many objects and some objects enclose other objects. The following heuristic was adopted to help solve the problem: *if an entity provides information and functions as a unit then it is classed as an object*. Thus a menu bar containing a number of different menu options is a single object whereas iconic buttons grouped together are each individual objects. The distinction here is a fine one and emphasises the importance of how items are grouped. It is not necessarily the case that individual users will apply the same heuristic as ourselves; in fact, it seems more likely that users of different experience will see objects in different ways. And if users do group items, should object boundaries be drawn around the group that the user sees rather than the individual items? The answer to this question is unclear. Bonsieppe tended to group items and treated titles and blocks of text as single objects. It is certainly possible that experienced users may see less objects on a screen than do inexperienced users.

In this paper we will evaluate the validity of the Bonsieppe technique by comparing values of the complexity C for GUI application screens, determined using equation (1), with users' perceptions of the "goodness" of the screen designs.

System and Distribution Complexity

In a previous study [5], values of system complexity C_S and distribution complexity C_D were determined for thirteen Microsoft Windows application screens, using the grouping heuristic described above. The results are shown in Table 1. The complexity C as determined by equation (1) is an additive quantity, thus enabling the total complexity of each screen to be computed by summing the system and distribution complexities. Thus in Table 1 the total complexity C is given by $C = C_S + C_D$. The complexity per object C_O is also computed and is given by $C_O = C/N$. It is seen that there is a large variation in complexity figures for the thirteen different displays, with the complexity of the most complex display screen (from Rockford) being some 66 times greater than the complexity of the least complex screen (from Microsoft Recorder).

Application	Obj. no. N	System comp. C_S	Dist. comp. C_D	Total $C= C_S+C_D$	$C_O = C/N$
MSRecorder	5	10.46	13.22	23.68	4.74
MSCalendar	17	76.59	101.28	177.87	7.54
Arachnid	60	96.22	388.89	485.11	8.09
MSCardfile	11	36.82	53.35	90.17	8.20
STW	23	50.75	122.60	173.35	8.60
Chartist	31	85.67	199.94	285.61	9.21
MSSolitaire	14	64.24	69.44	133.68	9.55
MSObjectPackager	23	89.73	143.55	233.28	10.14
ObjectVision	20	57.19	114.82	172.01	10.46
MSPaintbrush	61	239.61	467.89	707.50	11.60
MSWord	74	305.86	591.79	897.65	12.13
MSExcel	79	312.55	656.06	968.61	12.26
Rockford	104	582.42	989.56	1571.98	15.12

Table 1: Comparison of thirteen different Windows application screens in order of total complexity per object

The correlations between the number of objects N, the system complexity C_S and the distribution complexity C_D are high, as indicated by the coefficients in Table 2.

	N	C_S	C_D
N	1.00	0.93	0.99
C_S	0.93	1.00	0.97
C_D	0.99	0.97	1.00

Table 2: Correlation coefficients for combinations of N, C_S and C_D.

The strongest correlation is between C_S and C_D and a graph showing the relationship between these two variables is shown in Figure 1. It is seen that this relationship is approximately linear, with the slope of the line being $m \approx 0.52$. A graph of system complexity and distribution complexity versus number of objects is shown in Figure 2. These relationships are not linear and indicate that the rate of complexity increases with increasing number of objects.

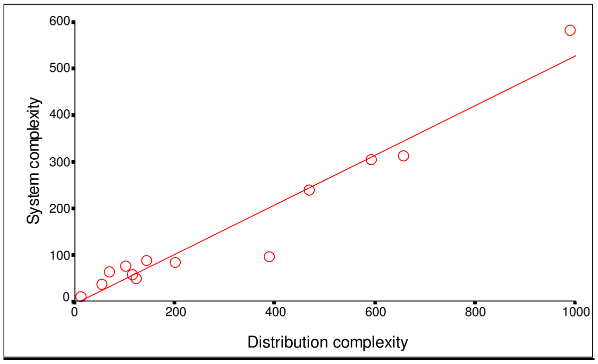

Figure 1: system complexity versus distribution complexity for 13 Windows application screens

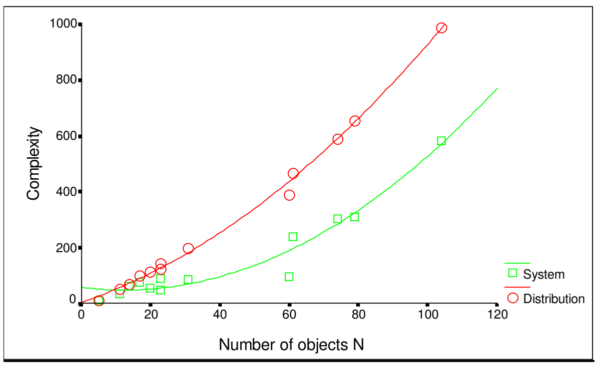

Figure 2: comparison of different measures of complexity

The data in Table 1 provide a theoretical means of ranking screens in order of complexity. However, the parameters C and C_O give rise to slightly different rankings (see Table 3 and Figure 3), with a Spearman correlation coefficient between the rankings of $r_s = 0.77$ at a significance of 0.002. The largest differences between these rankings (4 or more places) are due to screens from 3 applications: Microsoft Solitaire, Arachnid and ObjectVision. Arachnid (a card solitaire game with an appearance similar to MS Solitaire) becomes relatively less complex when C_O as opposed to C is used as the metric, moving from rank 9 to rank 3. Inspection shows that this change is due to the large number of identical and well ordered objects on the Arachnid screen. In contrast, due to a smaller number of objects less well ordered, both MS Solitaire and ObjectVision become relatively more complex when C_O is used instead of C. The sample size is unfortunately too small to make any generalisations here: it is therefore unclear from theoretical considerations alone as to which metric provides the better representation of complexity as seen by the user.

EXPERIMENTAL METHOD

Both Bonsieppe and Tullis have indicated that designs with high values of C are less desirable than designs with low values of C; this would also intuitively seem to be the case. On this basis, it would be expected that if users were asked to rank application screens in order of "goodness" of their design, then the ranking would be similar to that given in Table 1, i.e. Microsoft Recorder would be considered to be the best design and Rockford the worst. In practice, however, a user's perceptions of what constitutes a good or bad screen design will also be influenced by other factors; such factors might include the user's experience in computing and the user's knowledge of the application in question.

A survey was therefore carried out to determine whether Bonsieppe's technique would provide a predictive measure for users' ranking of different designs. Subjects were recruited from the local campus (both students and staff) and from off-campus. All subjects were volunteers and no rewards were offered. The survey took between 5 and 10 minutes to conduct.

A grey-scale 300dpi laser print was made of each screen and inserted in a plastic envelope. An extra copy of the Microsoft Word screen was also made for a pilot investigation of users' perception of grouping.

Each subject was tested individually. The subjects were asked to provide details concerning their:

- age
- sex
- occupation
- number of years of computer experience
- time spent per week using a computer
- expertise with a computer
- familiarity with Microsoft Windows

They were then asked to sort the screen prints from best design to worst design, with no ties. No attempt was made to define what was meant by "goodness" of design, this interpretation being left up to the subject.

Finally each subject was shown the Microsoft Word print for 5 seconds and asked to estimate the number of objects.

RESULTS

There was found to be a significant agreement in screen rankings between all thirty subjects, with Kendall's coefficient of concordance giving $W = 0.25$ and $\chi^2 = 91.1$ at a significance level of < 0.00005. The results indicate that, despite the lack of guidelines, subjects had a common interpretation of "goodness" of design. However, the distribution of the results was unexpected. The least complex screen in terms of either C or C_O is from MS Recorder. This screen was ranked as being the second *worst* design by 12 out of the 30 subjects. The most complex screen in terms of either C or C_O is from Rockford. This screen was actually ranked as being the *best* design by 4 subjects, although 9 other subjects ranked it as the worst.

The rankings by user perception are compared with the rankings by complexity in Table 3 and in Figure 3. Whilst the rankings by C_O show some positive agreement with the rankings by C, it is seen that there is lack of such agreement between either of these rankings and the rankings by user perception. In fact a Spearman correlation between ranking by perception and ranking by C gives a *negative* coefficient of r_s = -.52 at a significance level of 0.07, and a Spearman correlation between ranking by perception and ranking by C_O gives a negative coefficient of r_s = -.47 at a significance level of 0.11. Both of these correlations indicate that users show a greater preference for the *more complex* screens.

Application	ID	Mean Perceived Rank	Rank by total C	Rank by C_O
MS Paintbrush	1	4.4	10	10
MS Excel	2	4.5	12	12
MS Word	3	5.2	11	11
MS Solitaire	4	5.5	3	7
STW	5	6.0	6	5
Arachnid	6	6.1	9	3
ObjectVision	7	6.8	5	9
Chartist	8	7.0	8	6
MS Calendar	9	7.9	4	2
Rockford	10	8.1	13	13
MS ObjectPackager	11	9.5	7	8
MS Recorder	12	9.6	1	1
MS Cardfile	13	10.3	2	4

Table 3: Expected ranks compared to mean ranks

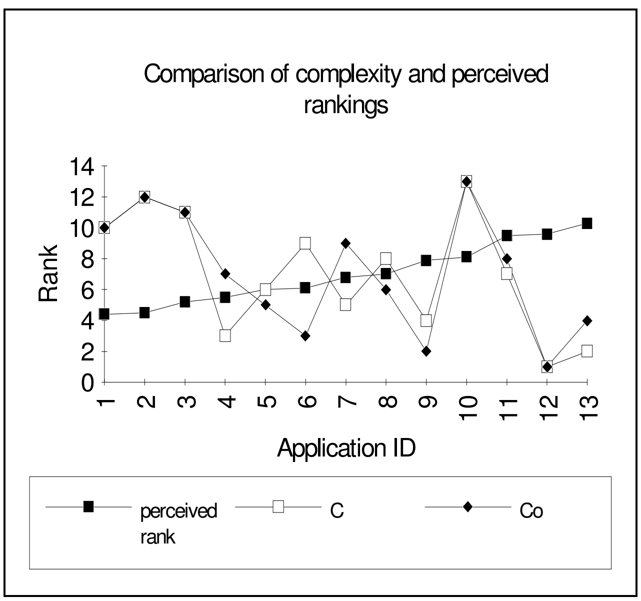

Figure 3: Mean ranks compared to expected ranks

Subjects are broken down by category in Table 4 and the Spearman correlation coefficient between the pairs of rankings in each category is given. Differences in sex, age and occupation make little difference to the rankings of the screens. The group showing the most disagreement in rankings is the "time spent on computers" category. The subjects who were expert, familiar with windows and spent more than 10 hours with a computer were classed as experienced users; this category shows a moderate agreement with the rankings from those who are inexperienced.

Category	N		N	r_s	Sig
Male	22	Female	8	.93	.000
Aged >40	14	Aged < 40	16	.83	.000
Student	11	Non-student	19	.76	.003
Expert	20	Novice	10	.72	.006
Familiar	20	Unfamiliar	10	.54	.055
> 10 hours	22	< 10 hours	8	.32	.284
Experienced	19	Inexperienced	11	.62	.025

Table 4: Spearman correlation coefficients for different categories of users

Table 5 gives the Spearman correlation coefficients between the C rankings and the least compatible group (>10 and <10 hours) and also the correlation between the C rankings and the experienced and inexperienced users. It is seen that those who use a computer for less than 10 hours a week have a significant negative correlation with the complexity rankings, whereas those who are more intensive computer users show a correlation closer to chance, though still negative. Similarly inexperienced users show a significant negative correlation.

Category		r_s	Sig
Ranked by C	All subjects	-.52	.071
Ranked by C	.> 10 hours	-.32	.279
Ranked by C	< 10 hours	-.79	.001
Ranked by C	Experienced	-.32	.279
Ranked by C	Inexperienced	-.72	.005

Table 5: Spearman correlation coefficients for experienced and intensive users

Subjects estimated that there were between 6 and 80 objects on the Word screen with an mean of 28. All but one subject underestimated the number of objects and that person indicated that he was a professional estimator. Most subjects felt that there were between 20 and 30 objects. There was no significant difference between the means for experienced and inexperienced subjects using a 2 sample t-test (t = 0.32, 0.05 sig.) and similarly there was no difference for time spent on a computer (t = 0.1, 0.05 sig.).

DISCUSSION

The expectation based upon the work of Tullis and Bonsieppe was that good layout design strives to be simple. This expectation was not borne out by our results; in fact a number of applications, including Microsoft Word and Excel, received rankings opposite to that expected. This suggests that users prefer more complex layouts. Inexperienced users are inclined to favour the screen designs that score highly on complexity and experienced users also show some preference for more complex screens. It is clear that in general users preferred more complex screen designs to simple ones (although experienced users did complain about the cluttered look of the Rockford screen). Even the three subjects who did no computing preferred the more complex screens.

There are clearly a number of problems with comparing screen designs for different applications. Some users reported being more familiar with certain designs and judged them better on the basis of familiarity, suggesting that screens may be judged to be "good" because users can map them to what they know the applications can do. However, as we have seen, the results show a high degree of correlation in screen rankings between all thirty subjects ($\chi^2 = 91.1$, $\alpha = 0.00005$), indicating that familiarity with the screen was not a major factor in the ranking process, despite the statements of the subjects familiar with the applications.

Some subjects expressed the view that the more objects on the screen the more functional the application (a large number of icons was a popular preference). This view is supported by the higher negative correlation between user rankings and C than between user rankings and C_O, since C as a complexity measure is a more direct function of the number of objects on the screen than is C_O.

It is possible that some results were affected by the use of paper models of the computer screens. However, subjects familiar with computers would be able to visualise the real screens and therefore it is believed that paper instead of screens would not have significantly altered the results.

Subjects were asked to estimate the number of objects in order to ascertain if there were differences in the way novices and experts grouped objects on the screen. The results showed no differences in the means for experience and time spent on computers. The only useful information the results provided was that users underestimate the number of objects on a screen. This is a complex area and we feel that further investigation is warranted.

Further Research

Further research is required to investigate users' preference for more complex screens and to determine if complexity theory has any value for comparing different designs of screens for the same application.

CONCLUSIONS

It appears that layout complexity is not an important issue for users deciding on the quality of a screen design. The most reasonable explanation is that screen designers already take into account layout complexity as design rules do indicate that screens should be laid out on a grid [10]. However, there is scope to investigate the effect layout complexity has on different designs for the same application.

The most interesting result is the degree to which people like complex interfaces. At first glance this is counter-intuitive but further thought indicates that people usually do tend to judge a tool by its perceived functionality. Manufacturers of low end sound systems have known this for many years; providing plenty of knobs, dials and readouts to impress the buyer. On the basis of this research it would be a good idea for interface designers not to open an application with a simple interface. Users like to see lots of goodies.

REFERENCES

[1] Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois.

[2] Bonsiepe, G. A. (1968). A method of quantifying order in typographic design. *Journal of Typographic Research*, *2*, 203-220.

[3] Tullis, T.S. (1983). The formatting of alphanumeric displays: A review and analysis. *Human Factors,* 25(6), 657-682.

[4] Tullis, T. S. (1988). A system for evaluating screen formats: research and application. *Advances in Human-Computer Interaction*, 214-286.

[5] Comber, T. (1994) A formal method for evaluating GUI screens. *CIS Working Paper 94/1, Centre for Computing and Mathematics, Southern Cross University.*

[6] Maltby, J. R. (1994) A method to quantify task complexity in WIMP interfaces. *CIS Working Paper 94/2, Centre for Computing and Mathematics, Southern Cross University.*

[7] Shneiderman, B. (1992) *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (2nd ed.), Reading, MA: Addison-Wesley.

[8] Hix, D. and Hartson, H. R. (1993) *Developing User Interfaces: Ensuring Usability through Product and Process*, New York: John Wiley & Sons.

[9] Gallitz, W. O. (1989*) Handbook of Screen Format Design*, Wellesley, MA: QED Information Sciences.

[10] Marcus, A. (1992) *Graphic Design for Electronic Documents and User Interfaces*, ACM Press.