EFFECTIVENESS AND PREFERENCES OF ANTHROPOMORPHIC FEEDBACK IN A STATISTICS CONTEXT

Dr Pietro Murano, Nooralisa Mohd Tuah
School of Computing, Science and Engineering, University of Salford, Gt. Manchester, M5 4WT, UK
p.murano@salford.ac.uk, a3lise@yahoo.com

Keywords: Usability, evaluation, anthropomorphism, interface feedback.

Abstract: This paper describes an experiment and its results concerning research that has been going on for a number of years in the area of anthropomorphic user interface feedback. The main aims of the research have been to examine the effectiveness and user satisfaction of anthropomorphic feedback. The results are useful to all user interface designers wishing to improve the usability of their products. There is still disagreement in the research community concerning the usefulness of anthropomorphism at the user interface. This research is contributing knowledge to the aim of discovering whether such approaches are effective and liked by users. The experiment described in this paper, concerns the context of statistics tutoring/revision, which is part of the domain of software for in-depth learning. Anthropomorphic feedback was compared against an equivalent non-anthropomorphic feedback. The results indicated the anthropomorphic feedback to be preferred by users for some factors. However the evidence for effectiveness was inconclusive.

1 INTRODUCTION

The usability of a system is sometimes neglected by software developers. However the usability of a software system and particularly the user interface is paramount to a system being successful in terms of user satisfaction, efficiency and success as a commercial product.

The aim of this research is to further knowledge concerning which types of feedback are effective and liked by users, particularly the authors are investigating the appropriateness of anthropomorphism at the user interface. Anthropomorphism at the user interface usually involves some part of the user interface, taking on some human quality (De Angeli, Johnson, and Coventry, 2001), e.g. a synthetic character acting as an assistant to users. Video of a human may also be viewed as being anthropomorphic (Bengtsson, Burgoon, Cederberg, Bonito and Lundeberg, 1999). Lastly various commercial objects have also been anthropomorphised (DiSalvo and Gemperle, 2003).

Anthropomorphic feedback has been researched for several years by various researchers around the world. Despite a reasonable amount of research being conducted in this area, there is still disagreement amongst researchers regarding the usefulness of anthropomorphism at the user interface. Some studies have shown anthropomorphic feedback to be more effective and preferred by users and in some other studies the converse has been shown. This kind of disparity can also be seen in the work of Murano and his collaborators (see Murano, 2002a, 2002b, 2003, 2005, Murano, Gee and Holt, 2007 and Murano, Ede and Holt, 2008).

An early study conducted by one of the authors (Murano, 2002b, 2005), concerned the context of English as a foreign language pronunciation. This was in the domain of software for in-depth learning. In this study, anthropomorphic and non-anthropomorphic feedback types were compared. Participants were given a series of pronunciation tasks and they were scored according to their self-correction success, based on automated feedback

given to them during the experimental sessions. The feedback was varied as being either anthropomorphic or non-anthropomorphic. The data collected for this experiment showed with significant results that the anthropomorphic feedback was more effective and preferred by users.

However in another study by Hongpaisanwiwat and Lewis (2003) in the tutoring type context, the authors investigated animated characters and voices in the context of graphics tutoring. They specifically dealt with participants' understanding and concentration in learning in relation to a set of learning materials. Their experiment had three conditions. These were an animated character, a 'pointing finger' and no character. Further, within these conditions they used synthetic or human voices. The authors also factored into the investigation some human personality types.

The general results of the experiment were that there was no significant difference in understanding for participants using the animated character. No significance was found with the synthetic character condition or the type of voices used in relation to what the participants remembered. An interaction effect was observed showing the animated character with synthetic voice condition to be better at helping participants remember relevant aspects of the learning materials, compared with the animated character with human voice. No significance or interaction effects were observed for the personality type groups, type of voice used and the animated character condition. Furthermore, the personality type, type of voice and animated character did not have any main effect on the participants remembering items which had been specifically emphasised. Further interaction effects were observed showing that the amount of emphasised items remembered by participants using the 'pointing finger' and no character conditions with a human voice was greater than participants using the synthetic voice. Also, participants remembered more in the animated character with synthetic voice condition compared with the animated character with human voice condition. There were no significant differences concerning participants' subjective opinions of the learning materials.

An interesting study was also conducted by Moundridou and Virvou (2002) in a software algebra tutor context. They examined the effects of using an anthropomorphic 'conversational' agent. Two conditions were tested in an experiment. The first condition had a synthetic face and accompanying voice. The second condition was the same as the first condition, but had the agent removed and was replaced by textual messages. The information presented was equal under both conditions.

The main findings of the experiment were that the time taken for the algebra tasks was not significantly different for the two groups. Questionnaire responses to do with participant attitudes towards the experienced user interface, showed significant results in favour of the anthropomorphic agent. Participants tended to enjoy the system more, finding it more useful and less difficult to use. Furthermore in relation to a post-test administered to the participants, there were no statistically significant results to suggest that the anthropomorphic agent helped participants to complete the test in a faster time with better overall results.

The remainder of this paper has two further main sections. Section 2 describes the conducted experiment, along with the main results and experimental conclusions. Section 3 discusses overall conclusions along with some proposals for further work.

2 STATISTICS EXPERIMENT

2.1 Aims

The aim of this experiment was to gather data concerning the effectiveness (i.e. errors and task time) and user satisfaction of different feedback types in the context of software aiding users to revise or remind themselves of how to carry out statistical procedures without the use of a statistical application. This could potentially help students either coming back to use statistics after some time and therefore requiring some reminders, or it could help students studying a basic course in statistics.

Two feedback types were tested – anthropomorphic and non-anthropomorphic. The anthropomorphic feedback was in the form of an animated character (Media Semantics, 2009) uttering the required instructional content feedback, while the non-anthropomorphic feedback consisted of textual content.

2.2 Users

- 24 participants, with a university education, took part in the experiment.
- All the participants in the study were in the 21-35 age range.
- A combination of male and female participants were used.
- All the participants had some knowledge of basic statistics. However all participants had not used statistics for some time and

had been taught the subject some time in the past.

- All participants had a basic knowledge of using a computer for general tasks.

2.3 Design

A between users design was used, where all 24 participants were randomly assigned to one of the two conditions being tested. A between users design was chosen because this would avoid learning effects in the carrying out of the tasks.

2.4 Variables

The independent variable was the type of feedback (instructions uttered by the animated character and textual instructions) and the type of tasks, consisting of viewing some instructional material and then taking part in a quiz.

The dependent variables were the participants' performance in carrying out the tasks and their subjective opinions.

The dependent measures were that the performance was measured by task completion success, i.e. whether a participant completed quiz questions correctly and the time taken to complete a quiz question. The subjective opinions were measured using a post-experiment questionnaire designed by the authors of this paper. The questionnaire had various questions relating to the general user interface and the participants' experience of using the prototype system.

2.5 Apparatus and Materials

The experiment was conducted using two rooms. The first room was used as a waiting room, while the second room housed the computers with the installed prototype system.

The second room contained four high end laptops with similar specifications. Each had the prototype system installed.

Headphones were used for the anthropomorphic condition, so that the same venue could be used for up to four participants at a time. The headphones allowed each participant in the anthropomorphic condition to listen to the verbal utterances without disturbing other participants.

Each participant was given a sheet of paper and a pen for note taking and calculation purposes.

Two questionnaires were designed and used in this experiment. A pre-experiment questionnaire was used for recruitment purposes and a post-experiment questionnaire was used for eliciting subjective opinions. The pre-experiment questionnaire contained some basic personal questions, e.g. age group etc. There were also questions about computer experience and statistics experience. The post-experiment questionnaire contained questions, using Likert type scales, eliciting opinions on the user interface and its components, the instructional material, the quiz and the participants' feelings during the interaction.

2.6 Procedure and Tasks

The first step was to recruit suitable participants who had some basic computer usage experience and some knowledge of basic statistics. However it was a requirement that the sample used should not have been using statistics recently and had not been learning about statistics recently. This was because the prototype developed was specifically dealing with the context of giving a user a reminder of how to calculate certain statistical tests and/or help someone learning some basic statistics. The recruitment was achieved by using an appropriate pre-experiment questionnaire.

The pre-experiment questionnaires were completed in room 1 and assuming the volunteers had the required profile, they were then asked to move to room 2 containing the laptops. This was done in clusters of four, since there were four laptops. Then the participants were given a brief verbal overview of the experiment and some basic instructions regarding how the experiment would be done. At this point each participant was given a sheet of paper with some details of the system and the details of the tasks they would do.

At all times during the experiment, three experiment assistants were present to observe users and provide help as required (although no help that would aid in completing the tasks was given). Prior to the experiments taking place, a meeting was held with all experiment assistants so that during the experiment sessions, consistency could be maintained between the assistants. The main consistency issue involved treating participants in the same manner regardless of experimental assistant.

The next step involved showing and explaining to participants how the system worked. Following this, the experiment began by participants listening to or reading (depending on experimental condition) an introductory part of the prototype. This then led to the actual tutorial section, which gave some instruction on some basic statistical 'tests'. When this was completed participants would go to the quiz section, where they would attempt some similar statistical problems to those discussed in the tutorial.

The system recorded the time to complete each quiz question and if the solution provided was

correct. Participants were also observed and any issues were recorded on an observation protocol.

At the end of the tasks, participants completed a post-experiment questionnaire which dealt with subjective opinions. When this was completed, participants were debriefed and then lunch was provided for them as a reward for their participation.

Overall each participant spent approximately 30-40 minutes on the tasks. The actual tasks in the experiment involved listening to instructional material about calculating the mean, median, mode, range, standard deviation and variance. Then quiz questions had to be answered for each of these areas of concern. Only one attempt per question was allowed for all quiz questions.

2.7 Results

The data were analysed in terms of their distributions, particularly the means and standard deviations. Significance testing was carried out by means of t-tests for between users designs. Where a significant result was observed, relevant tables are shown below. However for brevity, no tables are provided in instances of insignificant results.

Firstly the distribution tables are provided below (tables 1 -5) for the statistically significant data. Overall these show low standard deviations, indicating consistency in the participants' scoring.

Table 1: Helpfulness of Screen Appearance

Mean	4.58
Std Dev	0.50
Std Err Mean	0.10
upper 95% Mean	4.80
lower 95% Mean	4.37
N	24

Table 2: Precise Instructions

Mean	4.5
Std Dev	0.51
Std Err Mean	0.10
upper 95% Mean	4.72
lower 95% Mean	4.28
N	24

Table 3: Helpful Text Formatting

Mean	4.71
Std Dev	0.46
Std Err Mean	0.09
upper 95% Mean	4.90
lower 95% Mean	4.51
N	24

Table 4: Specific Quiz Questions

Mean	4.46
Std Dev	0.51
Std Err Mean	0.10
upper 95% Mean	4.67
lower 95% Mean	4.24
N	24

Table 5: Will Use Info Elsewhere Agree/Disagree

Mean	4.83
Std Dev	0.38
Std Err Mean	0.08
upper 95% Mean	4.99
lower 95% Mean	4.67
N	24

For effectiveness issues, the times taken for quiz questions and the rate of correct solutions provided for the quiz questions were each analysed by means of a t-test. These did not reveal any significance.

All of the responses provided by the participants through the post-experiment questionnaire were also analysed by means of t-tests. The post-experiment questionnaire contained questions, using Likert type scales, eliciting opinions on the user interface and its components, the instructional material, the quiz and the participants' feelings during the interaction. The scales used, ranged from one to five. In most cases (and in all cases for the tables below) one reflected the most negative opinion and five reflected the most positive opinion. Four questions concerning the participants' feelings during the interaction used the one score to reflect a positive response (not shown below due to insignificant results).

For the variables 'helpfulness of screen appearance' and 'group', the anthropomorphic feedback is significantly ($p<0.01$) rated as being more helpful than the non-anthropomorphic feedback. The t-observed is 2.76**. This can be seen in Table 6 below:

Table 6: Helpfulness of Screen appearance, Anth (Anthropomorphic) vs Non-Anth (Non-Anthropomorphic)

Difference	0.50	t Ratio	2.76
Std Err Dif	0.18	DF	20.89
Upper CL Dif	0.88	Prob > |t|	0.01
Lower CL Dif	0.12	Prob > t	0.01
Confidence	0.95	Prob < t	0.99

For the variables 'precise instructions' and 'group', the anthropomorphic condition is significantly ($p<0.01$) rated as being more precise than the non-anthropomorphic condition. The t-observed is 2.71**. This can be seen in Table 7 below:

Table 7: Precise Instructions, Anth vs Non-Anth

Difference	0.50	t Ratio	2.71
Std Err Dif	0.18	DF	22
Upper CL Dif	0.88	Prob > |t|	0.01
Lower CL Dif	0.12	Prob > t	0.01
Confidence	0.95	Prob < t	0.99

For the variables 'text formatting' and 'group', the text formatting in the non-anthropomorphic condition is significantly ($p<0.05$) rated as being better than the text formatting in the anthropomorphic condition. The t-observed is 2.42*. This can be seen in Table 8 below:

Table 8: Helpful Text Formatting, Anth vs Non-Anth

Difference	-0.42	t Ratio	-2.42
Std Err Dif	0.17	DF	17.15
Upper CL Dif	-0.05	Prob > |t|	0.03
Lower CL Dif	-0.78	Prob > t	0.99
Confidence	0.95	Prob < t	0.01

For the variables 'specific quiz questions' and 'group', the quiz questions in the anthropomorphic condition are significantly ($p<0.05$) rated as being more specific in nature than the same quiz questions in the non-anthropomorphic condition. The t-observed is 2.16*. This can be seen in Table 9 below:

Table 9: Specific Quiz Questions, Anth vs Non-Anth

Difference	0.42	t Ratio	2.16
Std Err Dif	0.19	DF	21.84
Upper CL Dif	0.82	Prob > |t|	0.04
Lower CL Dif	0.01	Prob > t	0.02
Confidence	0.95	Prob < t	0.98

For the variables 'will use info elsewhere' and 'group', the participants in the anthropomorphic condition stated they would more likely use the information they had viewed somewhere else (i.e. on some other occasion and/or setting) compared with the participants in the non-anthropomorphic condition. The differences are significant ($p<0.05$), the t-observed is 2.35*. This can be seen in Table 10 below:

Table 10: Will Use Info Elsewhere - Agree/Disagree, Anth vs Non-Anth

Difference	0.33	t Ratio	2.35
Std Err Dif	0.14	DF	11
Upper CL Dif	0.65	Prob > |t|	0.04
Lower CL Dif	0.02	Prob > t	0.02
Confidence	0.95	Prob < t	0.98

2.8 Discussion of Results

Overall there were no significant results for performance, where performance was measured by examining the correctness of quiz questions and the time taken. The reason for this outcome could be that the tutorial material was not advanced enough to allow differences to emerge with respect to the different feedback types. However, one potentially positive aspect of this result is that it gives some confidence that the sample chosen which was randomly assigned to the two different groups was approximately equivalent in nature. This had been one of the aims in the recruitment procedure. A further aspect that could have affected the lack of differences emerging in performance was that the sample size should have ideally been larger.

For the subjective questions answered by participants, where there was significance observed in the results, the significance indicated the anthropomorphic feedback to be mainly preferred over the non-anthropomorphic feedback. This was the case for helpful screen appearance, precise screen instructions, specific quiz questions and the likelihood of using the viewed information on some other occasion. One exception was for the text formatting, where the non-anthropomorphic feedback was significantly rated to be better. The reason for this single significant result for the non-anthropomorphic feedback is unclear. However one aspect that could explain the result, concerns the fact that the non-anthropomorphic condition used text to explain the concepts of the statistical material. The anthropomorphic condition used a character to utter or speak the same material resulting in much less text appearing on the screen. The anthropomorphic condition only had text on the screen to show the formulae used and to label some interface elements.

3 CONCLUSIONS AND FUTURE RECOMMENDATIONS

Although the outcomes of this experiment had few significant results for subjective preferences and no significant results for performance, the preferences do appear to lie in the anthropomorphic region. However, the authors of this paper conclude that there is still further work to do before one can make more concrete conclusions on whether anthropomorphism at the user interface is useful in terms of improving performance and subjective satisfaction in a statistical tutorial context.

However this work has been useful in two main ways. The first is that as far as we know no one has conducted an experiment of this sort investigating specifically the context of statistics

tutoring and anthropomorphism. The closest we have found was by Moundridou and Virvou (2002), which was in an algebra context (briefly reviewed in section 1). Their anthropomorphic feedback was used for giving instructions and feedback. However, in contrast, the work conducted by the authors used the anthropomorphic feedback (in one of the conditions) to also convey the instructional material. The second aspect of usefulness of this experiment has potentially shown the way forward for improving on this work. The work can be particularly improved by obtaining more participants, altering the tutorial material to be more advanced so that differences in the modes of presentation may emerge and by improving upon the procedure used during the experiment. The approach of using four laptops and only three experimental observers could have meant some bias being present in terms of inconsistent observation (despite efforts at controlling this) and some relevant occurrence not being noticed, due to the reality of at least one observer having to observe two individuals at once. A further aspect that could be improved is to link this work with perhaps some learning theories or other suitable theory of cognition/interaction.

ACKNOWLEDGEMENTS

Prof Sunil Vadera, Prof Tim Ritchings and the Universiti of Malaysia Sabah are acknowledged for their support of this research.

REFERENCES

Bengtsson, B, Burgoon, J. K, Cederberg, C, Bonito, J, and Lundeberg, M. (1999) The Impact of Anthropomorphic Interfaces on Influence, Understanding and Credibility. Proceedings of the *32nd Hawaii International Conference on System Sciences*, IEEE.

De Angeli, A, Johnson, G. I. and Coventry, L. (2001) The Unfriendly User: Exploring Social Reactions to Chatterbots, *Proceedings of the International Conference on Affective Human Factors Design*, Asean Academic Press.

Di Salvo, C. and Gemperle, F. (2003) From Seduction to Fulfillment: The Use of Anthropomorphic Form in Design, Proceedings of the Designing Pleasurable Products and Interfaces Conference, p. 67-72, c ACM.

Hongpaisanwiwat, C. and Lewis, M. (2003) The Effect of Animated Character in Multimedia Presentation: Attention and Comprehension, Proceedings of the 2003 IEEE International Conference on Systems, Man and Cybernetics.

Media Semantics (2009) http://www.mediasemantics.com.

Moundridou, M. and Virvou, M. (2002) Evaluating the Persona Effect of an Interface Agent in a Tutoring System. Journal of Computer Assisted Learning, 18, p. 253-261. Blackwell Science.

Murano, P, Ede, C. and Holt, P. O. (2008) Effectiveness and Preferences of Anthropomorphic User Interface Feedback in a PC Building Context and Cognitive Load. *10th International Conference on Enterprise Information Systems*, Barcelona, Spain, 12-16 June, 2000 - INSTICC.

Murano, P, Gee, A. and Holt, P. O. (2007) Anthropomorphic Vs Non-Anthropomorphic User Interface Feedback for Online Hotel Bookings, *9th International Conference on Enterprise Information Systems*, Funchal, Madeira, Portugal, 12-16 June 2007 - INSTICC.

Murano, P. (2005) Why Anthropomorphic User Interface Feedback Can be Effective and Preferred by Users, *7th International Conference on Enterprise Information Systems*, Miami, USA, 25-28 May 2005. INSTICC.

Murano, P. (2003) Anthropomorphic Vs Non-Anthropomorphic Software Interface Feedback for Online Factual Delivery, *7th International Conference on Information Visualisation*, London, England, 16-18 July 2003, IEEE.

Murano, P. (2002a) Anthropomorphic Vs Non-Anthropomorphic Software Interface Feedback for Online Systems Usage, *7th European Research Consortium for Informatics and Mathematics (ERCIM) Workshop - 'User Interfaces for All' - Special Theme: 'Universal Access'*. Paris (Chantilly), France 24,25 October 2002. Published in Lecture Notes in Computer Science - Springer.

Murano, P. (2002b) Effectiveness of Mapping Human-Oriented Information to Feedback From a Software Interface, *Proceedings of the 24th International Conference on Information Technology Interfaces*, Cavtat, Croatia, 24-27 June 2002.