

Stevens, R. D. and Edwards, A. D. N. (1996). *An approach to the evaluation of assistive technology*. in Proceedings of Assets '96, (Vancouver), ACM. pp. 64–71.

An Approach to the Evaluation of Assistive Technology

Robert D. Stevens and Alistair D N Edwards

Department of Computer Science
University of York
York
England
YO1 5DD

telephone: + 44 1904 432765
email: robert@minster.york.ac.uk

ABSTRACT

A valid criticism of many innovations in assistive technology is that they have not been evaluated. However, there are obstacles which make this form of technology difficult to evaluate according to conventional paradigms. The reasons behind this are discussed. A particular evaluation which endeavoured to circumvent those problems is described. The item evaluated was *Mathtalk*, a program to make mathematics accessible to blind people.

KEYWORDS: Evaluation, auditory interfaces, earcons, blind people, mathematics.

INTRODUCTION

Many assistive research and development projects are criticized for their lack of evaluation. It is understandable that people will ask if an innovation is worthwhile then that fact should be demonstrable in some objective manner. In practice, however, there are a number of factors which make such evaluations difficult.

There are a number of conditions which make conventional evaluation of assistive technology difficult. These will be discussed in this paper and then an example of an evaluation which attempted to side-step those problems is presented.

EVALUATION TECHNIQUES

The controlled test is part of classical scientific method. Essentially the test requires two matched groups, the experimental group and the control. The experimental group receives the 'treatment' under test, and the control group is treated identically *except* that it does not receive the treatment. At the end of the test the 'performance' of the two groups is tested and the hypothesis is that any difference is due to the treatment received by the experimental group.

This method has been developed and used successfully with simple systems. For instance, to test a crop fertilizer, the

treatment would be the application of the fertilizer. The control group would be sown in an otherwise identical environment (light, temperature *etc.*) and the performance would be measured as the yield of crop.

The first problem that arises in trying to apply the same methods to testing involving humans is that they are much more complex. It is more difficult to ensure that the two groups are identical and that they receive truly identical treatment. Unlike plants, people cannot be kept locked in a 'greenhouse', away from all external influences during treatment. Similarly, how can one say that members of the two groups are truly matched; a lifetime of experience and education can produce very different people.

Another component of the controlled test is in the treatment of the results. In the example of the fertilizer test, one would not take a single plant from each of the groups and measure its yield to say whether the fertilizer was effective. Even under the strictest controlled conditions we know that there will be variation. Even if the fertilizer works, it is quite likely that the smallest plant from the control group will be larger than the some of the plants from the treatment group. So, to measure just one from each group will probably give a misleading result, so instead we would measure samples from each group and apply a statistical test to their measurements. At its simplest, if the average (mean) yield from the test group is significantly greater than the average from the control, then we could conclude that the fertilizer works.

The problem with the statistical approach is that its reliability depends very much on the size of the sample. Basically, the larger the sample, the more reliable the results. If one measures 100 plants out of a crop of 1000 the average is more likely to be representative than if one were to measure just 10.

This paper thus has two objectives. One is to present the results of the evaluation of a particular development intended to be used by blind people¹. The second objective is to discuss some of the problems of evaluation and how they might be overcome, using this particular study as a model.

¹ The authors deliberately use the term 'blind people', rather than any of the more fashionable, 'politically correct' terms, to make it clear that the people in question are those with no useful vision. Should this offend any reader, we apologise.

PROBLEMS OF APPLYING THE PARADIGM

It has already been hinted above that one problem of using the controlled testing paradigm with people is that there is a great deal of uncontrollable variability between people. This is even more true when dealing with people with disabilities. In the example described below, the target population was blind people. Though all people so classed share the characteristic of having no useful sight, that is about all one can say. Factors such as their age at the onset of blindness, their education, other (possibly associated) conditions and so on mean that there is a large inherent variability.

Another problem with controlled testing occurs when the treatment to be tested is truly innovative, when there is no existing alternative to be used as the control. For instance, *Soundtrack* was the first Macintosh word processor which could be used by blind people, through the use of speech and non-speech sounds (Edwards, 1989). For its evaluation, it would hardly have been fair to sit some blind people in front of it and another group in front of an unadapted visual word processor and then ask them which they preferred! (See Edwards, 1987, for details of how the evaluation was in fact conducted).

Next there is the problem of testing on sufficient numbers to attain statistical reliability. The numbers of blind people in the population is relatively small. There is a range of practical problems associated with assembling a group of test participants. One must find people who are willing and able to give of their time. They or the tester may have to travel some distance.

The standard approach to 'smoothing out' any unavoidable heterogeneity of the test population (as mentioned above) is to increase the sample size, but that implies finding an even greater number of participants. This is likely to be difficult, or even impossible.

EVALUATION OF MATHTALK

Having set out the problems of applying the classical controlled testing paradigm to the evaluation of assistive technology, we will go on to describe an evaluation study we have carried out which has succeeded in overcoming most of the problems. First we describe the product to be evaluated and then the study which was carried out. Full details of Mathtalk and of the evaluation will appear in Stevens (1995).

Access to mathematics for blind people

Whereas mathematics is an intellectual activity – it is done 'in the head' – communication of mathematics is always performed in some written, visual form. Whenever mathematicians gather together there will be a pencil and paper or chalkboard handy. Furthermore, communication with ones self is important in performing mathematics. Written notations are used as a form of *external memory*, used to store intermediate results during calculations.

This simple mechanical problem of access means that mathematics can be a difficult topic for people with visual disabilities. While there is no reason to suppose that such people are any less able to perform mathematics, if they cannot communicate the material (including with themselves) their progress will be hampered. This was demonstrated in a survey (Bormans and Cahill, 1994) which showed that visually disabled

students performing mathematics were distracted by the *mechanical* problems of the necessary manipulations from the conceptual requirements of the tasks.

Blind students require access to mathematical material in a non-visual form. The *Mathtalk* program addresses that idea by providing auditory representations of algebra. A blind person can read (but not manipulate) algebra using Mathtalk.

There are three major components to the Mathtalk program. First it will read out algebraic expressions using synthetic speech but in a manner which is unambiguous. A major limitation of speech presentations is the lack of control; one cannot take in all the information in a long, complex expression if it is spoken all at once. So, the second component of Mathtalk is a browsing language which the user can employ to navigate around expressions and control the information flow. The third main component of Mathtalk addresses the problem that speech is a relatively slow medium. While the information in speech is complete, there are situations in which it is better to get an incomplete but quick *glance* at an expression. That is the role of the non-speech *algebra earcon* of Mathtalk.

Mathtalk uses the prosodic content of speech to resolve ambiguity (Stevens, Wright and Edwards, 1994; Stevens, Wright and Edwards, 1995). For instance, visually the equations

$$3x + 4 = 7 \quad (1)$$

and

$$3(x + 4) = 7 \quad (2)$$

are clearly distinct, and mathematically their meaning is very different. However in normal synthetic speech the presence of the parenthesis is likely to be lost. They can be explicitly spoken, but this lexical overhead tends to interfere with comprehension. Instead Mathtalk reads the expression as a human reader might, inserting pauses around the parenthesized sub-expression and speaking it with a lower pitch and increased speed.

The browsing language of Mathtalk enables the user to move through expressions eliciting a speech and non-speech representation of the material but in such a way that the user maintains *control* of the information flow (Edwards and Stevens, 1993; Stevens, Wright and Edwards, 1995). Commands have two components, an action and a target. Actions include **current**, **next** and **previous** and targets are mathematical entities such as **equation**, **term** and **item**. The commands are expressed as the initial letters of the commands, so that **current expression** would be entered as **ce**, for instance.

Algebra earcons are described in greater detail in Stevens, Brewster et al. (1994). As with all earcons, they use the rhythm, pitch and timbre of non-speech sounds to encode information. In the case of algebra earcons, the rhythm and pitch patterns are borrowed from those of speech. Timbres (musical instruments) are used to represent the different components of the expression.

By replacing mathematical objects in the expression with musical tones, that represent the class but not the instance of an object, we can give a high-level view of the structural nature of the expression without overloading the listener with the detail of the expression. Using a stylised form of the same prosodic cues in the spoken expression, the form of the audio glance matches

that of the spoken output and gives a more consistent presentation of an expression.

Non-speech sounds are also used to signal the beginning and end of components of expressions. For instance, in moving from left to right through an expression when a new complex component is encountered a rising, opening sound is heard and at the end of that component a falling terminus sound is heard. The same musical sounds used in the audio glance are reused in these terminus sounds giving a consistent feel to the interface and hopefully making it easier to learn.²

The evaluation described below was carried out as part of the development of Mathtalk, but the ideas behind Mathtalk have been adopted and extended within the *Maths* Project. The objective of this project is to build a workstation on which blind and visually impaired students will be able to not only read textual mathematical material, but will also be able to manipulate it. It will have a multi-modal interface, incorporating an enhanced visual display, speech input and output, non-speech sounds and braille.

Problems of evaluation

The evaluation of Mathtalk is subject to most of the problems of controlled testing listed above. It is a true innovation. That means that there is no one generalized alternative against which to test it. Currently blind mathematicians cope with a variety of representations including audio tapes, braille (using a variety of mathematical braille codes) and typesetting notations (such as Tex and Latex, Knuth, 1987; Lamport, 1988).

The Maths Workstation is aimed at upper-secondary-school-level (16-18 year-old) mathematics students. This was a deliberate choice because it means that a certain level of mathematical knowledge can be assumed in the users. It is recognized that ultimately the facilities must be usable by younger students because if they do not learn the early fundamentals of mathematics they will never attain the level of the more advanced material. However, it was decided that this was a more difficult challenge which could not be tackled until the easier one had been solved.

The very fact that accessing the material is difficult probably dissuades many visually disabled students from studying it any further than necessary. It is thus envisaged that (if successful) the Maths Workstation will create a new population of visually disabled mathematicians. In the meantime, however, the number of people qualified to test the workstation and its components are small. It was decided that it would not be possible to establish two groups for a controlled test. It would also be impossible to match participants; their backgrounds were simply too diverse: level of sight, aetiology of disability, level and form of education and so on.

Another problem with the controlled testing paradigm is that it requires the measurement of something. In the context of the Maths Workstation, what would one measure? To grade the answers of mathematical exercises would not be fair – are errors due to difficulties in the interface or to lack of mathematical

understanding in the participant? The evaluation should test the innovations of the user interface, not the mathematical ability of the participants.

The problems of recruiting a sufficient number of testers for statistical significance has been alluded to above. In this case the difficulties were exacerbated by the fact that mathematics is not a popular activity. No matter how important it may be, most people do not enjoy their exposure to it at school and avoid it thereafter as much as possible. Therefore in this case, we had to find people who were not only able to perform the

² Paper is an inappropriate medium for describing sounds. Readers who want to hear the sounds used can consult the World Wide Web pages at:
<http://dcpu1.cs.york.ac.uk:6666/hci/aig/alistair/maths.html>

No.	Mathtalk	Typeset	Latex 'Raw'
1	$y = 7x + 3$	$y = 19 - 3x$	<code>y = 19 - 3x</code>
2	$y = (x + 3)(x - 2)$	$(x + 3)(x - 3) = y$	<code>(x + 3) (x - 3) = y</code>
3	$y = \frac{1}{2}(x + 5)^2$	$y = \frac{1}{3}(x + 5)^2 - 7$	<code>y = \frac {1} {3} (x + 5)^2 - 7</code>
4	$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$	$p = \pm \frac{lx_1 + my_1 + n}{\sqrt{l^2 + m^2}}$	<code>p = \pm \frac {lx_1 + my_1 + n} {\sqrt {l^2 + m^2}}</code>

Table 1. Example expressions used in the evaluation. The *Latex*, *Typeset* column shows the expression while the 'Raw' column shows the text as it was represented in the Latex. Notice that the raw Latex was spaced in such a way as to facilitate reading.

testing required, but who were willing to do so.³

Design of the evaluation

Evaluation was an inherent part of the Mathtalk Project; it was not something simply tagged on at the end for the sake of credibility. Each component of the program was evaluated separately and aspects of the design modified in accordance with the results. In this way, the design and development was incremental and iterative. However, a point is inevitably reached when the whole system has to be put together and the question addressed as to whether it works.

Remember that Mathtalk provides only read access to mathematical material. The development of manipulation facilities is part of the Maths Project, for which a further evaluation is being carried out (Cahill, Linehan et al., 1995).

It was decided that the evaluation should have a control component, despite the problems of controlled testing, described above. There were three realistic possible controls: amanuensis (a human sighted reader responding to directions of the participant); audio tape recordings or use of a word processor. One way of reducing the numbers of test participants required is to not have two separate groups, but to have one group which tries both interfaces. This also gets around the problem of having to match the groups. However, it does incur possible spoiling of the results due to practice. Each participant will use one interface and then the other. Lessons learned with the first interface may transfer and affect performance and preferences for the second one (positively or negatively). The way to try to cancel this effect is to introduce half the participants to one interface first while the others start with the second interface, then the groups effectively swap.

The word processor option was chosen. Material was available in a machine-readable form in Latex notation and participants were regular users of a word processor / screen reader combination. The Latex option was the one found to be most commonly used by practising blind mathematicians. (In contrast, in a survey in Ireland *no* school students were found to use audio recordings of mathematics, see Cahill and McCarthy, 1994). The use of a human reader was rejected because it implied an unfair comparison. The user of the Mathtalk program will be an independent person who will not be able to rely on the mathematical knowledge and assistance of others (except in a formal teaching situation).

Cooperative evaluation is an alternative to the strict controlled experiment paradigm which is appropriate to the evaluation of human-computer interfaces (Wright and Monk, 1989). A modified version was used in this case. Blind participants were given tasks to perform under the two conditions. As they worked they were encouraged to think aloud. Performance of the tasks and recordings of the commands issued were used to give a measure of performance and an indication of the style of operation developed. Such quantitative measures do not tell the whole story, though and so subjective data was also collected.

The Task Load Index (TLX) was developed by NASA as a means of measuring the workload under test conditions (NASA, 1987; Hart and Staveland, 1988). It has successfully been used as part of the evaluation of multi-modal human-computer interfaces by Brewster (1994). To use it, test participants are asked to give numerical ratings on a scale of 0 to 20 for:

- mental demand,
- time pressure,
- effort expended,
- perceived performance and
- frustration experienced.

Other subjective data was collected as to the participants' preferences and other comments on the interfaces.

Procedure

Broadly the procedure was as follows. Participants were divided into two groups, WP and Mathtalk. Each individual was given approximately 30 minutes of training with their interface. They were then asked to perform a set of tasks using their interface and a set of supplied mathematical material. During and after the

³ This was in clear contrast to the evaluation carried out as part of the Guib project (Crispen and Petrie, 1993). There the objective was to make graphical user interfaces accessible to blind people and since most blind computer users have encountered problems accessing GUI software they were only too willing to try out something which might have a clear benefit to them.

exercises data was collected. Then the members of each group were moved to the other interface, the WP people were given the Mathtalk exercise and vice-versa. The word processing test were carried out using the WordPerfect word processor and IBM Screen Reader 2.

There were four participants. All were blind but varied in their age at the onset of blindness. They all had experience of using computers. They varied in their level of mathematical education but were all accustomed to using linear, programming-language-like mathematical notations (usually self-defined ones) in word processors.

Two sets of algebraic expressions were devised, which were judged by independent assessors to be of comparable syntactic complexity and mathematical 'difficulty'. One set was entered into the Mathtalk system while the other was translated into Latex. Examples are shown in Table 1. The Latex, Raw column in the table shows the unprocessed Latex as used in the Word processor condition. For example, Equation 3 would be read out as 'y equals backslash frac, open brace one, close brace, open brace three, close brace, left paren x, plus five, right paren, circumflex two, minus seven.'

A set of tasks was devised for testing under each condition. The tasks were similar, but adapted according to the facilities of the interface. Tasks involved navigation (e.g. 'Move to Expression 2; explore and describe this expression') and substitution with (simple) evaluation (e.g. 'What is the value of Expression 11 if x = 2?'). The experimenter was present during the testing and answered any questions posed by the participants. Audio recording were made to capture the dialogues and the participants' think-aloud protocols.

Data collected

After the exercises the TLX scores were recorded and then a structured interview was carried out based on a questionnaire which covered the following broad categories:

- 1 How good at computing would someone have to be to use the presentation to do the tasks?
- 2 General presentation of expressions.
- 3 Navigation and orientation.
- 4 Doing the tasks.

Examples of detailed questions (under categories 2 to 4 respectively) were as follows (full details are in Stevens, 1995):

- How easily could you tell different parts of the expression apart?
- What techniques did you use to move to a new term?
- In what ways were some question more difficult than others?

Participants were also asked to rate their level of preference of each of the two interfaces. On a 0 to twenty scale they were asked to score 0 if they much preferred the word processor over Mathtalk, or 20 if their preference was the other way around. A score of 10 would indicate that they had no preference.

Results

The approach to the tasks was very different in the two conditions in terms of the commands used. In the Word processor condition participants tended to move around

character-by-character (using cursor keys), whereas the Mathtalk users employed its browsing command to move by mathematical terms. This difference is reflected in the numbers of commands issued, as shown in Table 2.

The total times to complete the tasks were measured by timing the tape recordings. Time spent in dialogue with the experimenter were excluded.

Condition	Participants				Mean
	1	2	3	4	
Mathtalk	239	239	322	341	285
Word processor	642	661	617	549	617

Table 2. Numbers of commands issued under each of the conditions.

The times for the navigation tasks are shown in Table 3. A difference in the times is apparent, but this is not statistically significant, based on a paired sample two-tailed t-test ($t = -1.56$; $df = 3$; $p = 0.11$). The lack of significance is explained to some extent by the exceptionally fast performance by one of the participants in the word processor condition.

Times for the evaluation tasks are shown in Table 4. Here the differences are clearly non-significant ($t = -0.3113$; $df = 3$; $p = 0.39$). Two of the participants were faster in the Word processor condition.

Condition	Participants				Mean
	1	2	3	4	
Mathtalk	46	93	45	90	69
Word processor	85	139	63	73	90

Table 3. The average times (in seconds) for participants to complete each of the navigation tasks, along with the overall averages.

Condition	Participants				Mean
	1	2	3	4	
Mathtalk	99	105	67	99	93
Word processor	86	110	80	102	94

Table 4. The average time (in seconds) for participants to complete each of the evaluation tasks, along with the overall averages.

The timing results appear equivocal; they do not suggest a clear advantage for the Mathtalk condition. This may be partly due to the fact that the participants were experienced WordPerfect users, while they were essentially novices at using Mathtalk. Nevertheless there are positive conclusions from this part of the evaluation. Mathtalk has a complex (and unfamiliar) interface, so that for people to perform no more slowly than with WordPerfect / Latex is quite promising. It seems likely that with practice people could become much more efficient, as stated by one of the participants, 'I can see that once the commands have been learnt, this could be a very, very fast way to read expressions'.

It was also observed that under the Mathtalk condition the participants were more willing to explore. They would find the

answer to the current exercise but then use the facilities to discover more about the current expression. This meant that their recorded times were increased compared to the word processor condition, under which they concentrated on getting enough information to answer the question and then stopped.

The mean preference rating (on a scale where a score of 20 implies an absolute preference for Mathtalk) was 16. There were three extreme scores (17, 17 and 20). One participant indicated 'no preference'. He explained that this was because he was so used to the word processor style of working. In that case it was quite positive that he did not err more towards the zero end of the scale (but see the discussion below on *Subjects' motivation*).

Observations

In the word processor condition use of the cursor to move between expressions was very efficient. For instance, to move back five expressions, the user would simply press the up cursor key five times. Sometime there were problems, though, when the user became confused about where the end of the line was.

Under the word processor condition the participants were very consistent in not listening to the whole of an equation at a time. Instead they preferred to move straight into the equation, using the cursor controls to hear it character-by-character. In contrast, in the Mathtalk condition participants frequently used the **show expression** command. This causes the whole of the current expression to be spoken in detail. The current level command was also frequently used. This command speaks the expression, but hides complex objects by referring to only their type: $3(x+4)=7$ would be spoken as '3 times a quantity equals seven'. The algebra earcon is another facility which gives an overview of an expression for planning purposes. This was used frequently in the navigation tasks, but not in the evaluation tasks. The non-speech sounds representing the terminus of components were also well liked and widely used.

It was interesting to observe that users of the word processor frequently felt the need to mute the speech. This suggests that too much speech was being generated. In contrast, under the Mathtalk condition there was no demand for a mute facility. This implies that the control facilities built in to Mathtalk were of an appropriate level, that they produced just the right amount of (auditory) information.

Mental workload: TLX scores

Scores on the TLX scales are shown in Table 5. The overall mental workload was calculated from these figures as follows. For all the factors except Perceived Performance a high score (near to 20) represented a positive result. For example, a score of 20 for Mental Demand implied that the demand was very high. On the other hand, a score of 20 for Perceived Performance would imply that the user judged he or she had done well. For that reason the Perceived Performance scores were complemented (subtracted from 20) and an overall score calculated. Thus an overall score approaching 20 implies that the interface is difficult to use and that the results achieved are perceived as bad, whereas a score near zero means that the interface is easy to use with good results.

In the event the overall mean mental workload was calculated as 5.5 for the Mathtalk condition and 10.2 for the word processor condition. This difference is significant ($t = -2$, $df = 4$, $p = 0.04$).

The scores for Mental Demand are significantly different ($t = -7.52$; $df = 3$; $p = 0.005$). This is a very important result, suggesting that Mathtalk was well designed to meet the requirements of the task. Making the mathematical expressions more usable in terms of reducing mental demand involved in reading, releases more mental resources for performing the mathematical tasks.

The Effort Expended score differences were not statistically significant. Since the physical effort required for both interfaces was low this is more a measure of the mental effort involved. The Time Pressure scores were also not significantly different. This is not surprising since users were not given any time limits.

The identical scores for Perceived Performance demonstrates that the information necessary to perform the tasks was available under both conditions. However, the scores for Mental Demand suggest that this information was more easily accessed under the Mathtalk condition. The Frustration Experienced in the word processor condition is much higher than with Mathtalk, but this is not statistically significant ($t = -1.4$; $df = 3$; $p = 0.25$). This was due to the fact that one participant found the Mathtalk condition much more frustrating. He attributed this to a dislike of the style of keyboard commands (he had only used a portable computer keyboard and was unfamiliar with the full desk-top style of keyboard).

Factor	Word processor	Math-talk	Difference	% Difference
Mental Demand	14.8	7.0	7.8	210.7
Time Pressure	8.3	6.5	1.8	39.0
Effort Expended	9.8	3.5	6.3	30.0
Perceived Performance	12.3	12.3	0.0	0.0
Frustration Experienced	10.5	2.8	7.8	39.0

Table 5. The task load index (TLX) scores for the different conditions. The scores are in the range 0 to 20. For all the factors except Perceived Performance a high score (near to 20) represents a positive result.

SUBJECTS' MOTIVATION

In the section on *Problems of applying the paradigm* we discussed some of the limitations of the controlled testing paradigm when applied to the evaluation of assistive technology. There is another problem which may have affected the results in the evaluation presented here. As with some of the other limitations, the cause is a practical one, that of shortage of personnel. In this and other studies, the person doing the evaluation is the same person who developed the artefact being

tested. Test participants are usually volunteers and sometimes their over-riding motivation may be to *please the evaluator*.

This was evident in some of the questioning of the participants, who said things like ‘What do you want me to say?’ One way around this is to have the evaluation carried out by someone other than the developer of the artefact, someone who would not be identified by the participants as having an emotional attachment to it. This is often impractical, given the limitations on research funding. Another alternative is to deceive the participants, to make them believe that the objective of the testing is different from its true nature (for instance that someone else developed the artefact and that the objective is to demonstrate how bad it is). Clearly that alternative is very dubious ethically. It might also lead to a negative bias.

DISCUSSION

An evaluative study has been described which attempted to surmount the problems often experienced in evaluation of assistive technology. Many of the difficulties have been overcome so that we have an evaluation which we are confident gives a fair picture of the achievement. Nevertheless it has to be recognized that not all the limitations have been surmounted.

For instance, the problem of achieving statistically significant results because of small numbers of testers has been mentioned. In the case of the study described herein four testers were recruited and statistical tests were applied to the numerical results. The small numbers of testers may have affected the results where significance was not achieved, but by the same token too much should not be read into the results calculated to be significant. For instance, it has been noted above how the extreme view of a single participant could have a large effect on the results. Only by testing with large numbers would it be possible to say whether this reflects the outlook of something like 25% of the population – or whether it is peculiar response.

Subjective, questionnaire-based information elicitation is of some value, but the results must be treated carefully. The TLX provides a quantitative measure of users’ reaction to the artefact. It remains for the implementation of the Maths Workstation and its full evaluation to be completed before it will be possible to see whether those scores are true predictors of performance.

ACKNOWLEDGEMENTS

The Maths Project is funded by the European Union’s TIDE Initiative, project number TP1033. Most of the work described in this paper was carried out as part of a PhD project, funded by the UK Engineering and Physical Sciences Research Council, studentship number 91308897. Thanks are also due to Margaret Uffendell and the Royal National Institute for the Blind’s Vocational Training College for their assistance in the evaluation.

REFERENCES

Bormans, G. and Cahill, H. (1994). *Problem Analysis: A Formative Evaluation of the Mathematical and Computer Access Problems as Experienced by Visually Impaired Students*, Deliverable No. D1, The Tide Maths Project.

Brewster, S. A. (1994). *Providing a structured method for integrating non-speech audio into human-computer interfaces*. PhD Thesis, University of York, UK.

Cahill, H., Linehan, C., et al. (1995). Ensuring usability in Maths. in I. Placencia-Porrero and R. P. d. l. Bellacasa, *The European Context for Assistive Technology: Proceedings of the Second Tide Congress*, (Paris), IOS Press. pp. 66–69

Cahill, H. and McCarthy, J. (1994). Usability Analysis, Deliverable No. D3, The Tide Maths Project.

Crispen, K. and Petrie, H. (1993). Providing access to GUIs for blind people. in *Proceedings of the 19th Convention of the Audio Engineering Society*

Edwards, A. D. N. (1987). *Adapting user interfaces for visually disabled users*. unpublished PhD Thesis, Open University, UK.
Edwards, A. D. N. (1989). Soundtrack: An auditory interface for blind users. *Human Computer Interaction* 4(1): pp. 45-66.

Edwards, A. D. N. and Stevens, R. D. (1993). Mathematical representations: Graphs, curves and formulas. in D. Burger and J.-C. Sperandio (eds.) *Non-Visual Human-Computer Interactions: Prospects for the visually handicapped*. Paris, John Libbey Eurotext. Edition, pp. 181-194.

Hart, S. and Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. (in) P. Hancock and N. Meshkati (eds.) *Human mental workload*. Amsterdam, North Holland B.V., pp. 139-183.

Knuth, D. E. (1987). *The TeXbook*. Reading, Massachusetts: Addison-Wesley.

Lampert, L. (1988). *Latex: Users Guide and Reference Manual*. New York: Addison Wesley.

NASA (1987). *Task Load Index (NASA-TLX) v1.0 computerised version* NASA Ames Research Centre.

Stevens, R. (1995). *Principles for the design of auditory interfaces to present complex information to blind computer users*. DPhil Thesis to be submitted to the University of York, UK.

Stevens, R., Wright, P. and Edwards, A. D. N. (1995). Strategy and prosody in listening to algebra. in G. Allen, J. Wilkinson and P. Wright, (eds.) *Adjunct Proceedings of HCI’95: People and Computers*, (Huddersfield), British Computer Society. pp. 160–166

Stevens, R. D., Brewster, S. A., Wright, P. C. and Edwards, A. D. N. (1994). Providing an audio glance at algebra for blind readers. in G. Kramer and S. Smith (eds.), *Proceedings of ICAD’94*, (Santa Fe Institute, Santa Fe), Addison-Wesley. pp. 21-30

Stevens, R. D., Wright, P. C. and Edwards, A. D. N. (1994). Prosody improves a speech based interface. in D. England (ed.) *Ancillary Proceeding of HCI’94*, (Loughborough), British Computer Society.

Wright, P. C. and Monk, A. F. (1989). Evaluation for design. in A. Sutcliffe and L. Macaulay (eds.), *People and Computers V: Proceedings of HCI’89*, Cambridge University Press. pp. 345–358

