Validity and Cross-Validity in HCI Publications

Harold Thimbleby

Department of Computer Science, Swansea University, Wales
h.thimbleby@swansea.ac.uk

Abstract. Papers in HCI play different roles, whether to inspire, solve industrial problems or further the science of HCI. There is a potential conflict between the different views, and a danger that different forms of validity are assumed by author and reader — deliberately or accidentally.

This paper reviews some of the issues in this complex area and makes practical recommendations. In particular, the paper introduces the term "cross-validity" to help make explicit the issues, problems and means to tackle them.

1 Background

Errors in published scientific papers play different roles. Resolving an error may advance science, it may uncover fraud, or it may remain undetected and delay progress or it may (being undetected) cause inexplicable and apparently unavoidable problems. At another extreme, an inspiring paper may be no less inspiring despite manifest errors — researchers will be stimulated to sort out the errors and inaccuracies they wish to overcome.

Papers can be sound but incomplete; or, a common problem, the analysis correct, but the data flawed. There seem to be many ways for errors to creep in. Papers can be valid in at least three different senses: they may be objectively valid; they may appear valid; or they may be effective for the community (or some sub-community) of researchers. Philosophers may further argue that objective validity is unattainable in any case — there is no rational truth to be 'valid' about in a paper.

In HCI, we have different sources of confusion or possible confusion over types of validity:

- Many techniques in HCI are developed to be used on prototypes or approximations, particularly techniques intended for application in system evaluation (e.g., cognitive walkthrough). It is then a short step to do scientific research with prototypes and approximations instead of real, robust systems.
- Doing good HCI (or usability) involves a particular process, typically starting with something like task analysis, progressing through prototyping and implementation, then evaluation, then iteration. If any HCI publication must show evidence of this process to be valid, then some sorts of HCI may be being excluded. This is a particular problem with doctoral work in HCI, where examiners may expect a three year or longer research project to exhibit all features of the HCI process.

G. Doherty and A. Blandford (Eds.): DSVIS 2006, LNCS 4323, pp. 11–24, 2007.

- HCI is of course multidisciplinary. At worst, one discipline's validity is another's irrelevance. Computer scientists may do work ignoring psychology, and *vice versa*. Mathematicians may do work that cannot be implemented. And so on. Grudin covers these issues very well [9].
- Almost all work in HCI involves a complex system, one or more humans, a task and an environment. Few of these are ever fully defined in a paper, or understood by an author; in fact, the user and environment rarely *can be* defined, and the interactive system itself is rarely available for inspection. In short, work in HCI is based on approximations — that may compromise validity.
- Since a goal of HCI is usability, then it has been argued publications should be usable. If this is believed, then hard papers (e.g., containing mathematical theory) will be published less.
- Usability is the improvement of specific products, for instance in production, whereas HCI as the field of research, for instance refining the principles of design. However the words are defined, there is a conflict on what valid work is. For example, Wixon [20] claims very strongly that the current HCI literature fails — probably because he applies a usability perspective to papers that might not themselves claim to be usability but HCI. Usability professionals read and referee the HCI literature, and their standards, particularly concerning rigour, the significance of errors and handling errors is pervasive in the field. Whether that matters, and if so, from whose point of view, is a crucial point.
- More generally, HCI is a multidisciplinary subject, with disciplines drawn from a very wide variety of traditions. Notions of validity are especially hard to appreciate across disciplinary boundaries, because they are often implicit in the respective discipline's traditions. For example, a mathematician may not appreciate the difficulty in even identifying research questions in a soft approach; a social scientist may not appreciate the difficulty of programming correctly; and a programmer may not appreciate the nature of rigorous experimental methods with human subjects (let alone the ethical issues). A recent book in activity centred design writes, "Leont'ev (1981) created a formal structure [that is] less a representation of reality than a heuristic aid" [6]. To a mathematician this makes as little sense as modal logic must to an activity theorist; yet both can contribute to HCI, and will contribute more if we can find ways to bridge the disciplines — *whilst* remaining true to the disciplinary notions of validity.

Those are brief presentations of deep issues — that we will not resolve in this short paper! The point is to indicate the nature and depth of the problems. What the tensions represent is that there are many ways for author and reader of papers to have differing approaches to validity. Indeed, in HCI this tension seems inevitable. How can we reduce or resolve the tensions? How can we agree to differ where differing is appropriate? Some even argue (with the backing of usability experiments) that validity itself is not a valid notion in HCI [12].

We often have a naïve view of validity. "The scientific researcher writes objectively and is devoted to the pursuit of truth, regardless of pressures such as career progression, financial inducement, fame, or whatever." If we think or teach this simplistic view, then dealing with the reality of error in research will be even harder. Neither readers nor writers of papers will be prepared to guard against potential problems — because they do not expect them. Indeed, referees will not be prepared either, and poor papers will slip through. In turn, the next generation of students will read the poor papers and think that they set the standard to aspire to; thus the next researchers will likely work to lower standards than the ideals of the previous generation.

Errors may mislead readers of a paper and waste time as researchers attempt to reproduce work that has been inaccurately reported. Perhaps worst is when a community ignores minor errors, and standards drop. Perhaps nobody minds if science progresses faster because putting less effort into polishing papers for publication means they can be published faster, but in the long run lowering standards lowers publishing standards. Again: a vicious cycle ensues: poor publications are taken to define the standards of acceptable research, and worse work is then published to appear to be of that standard. New researchers do not learn or appreciate rigour.

The *appearance* of validity is a problem: researchers may read a paper, work from it, but ultimately be wasting their time — does it appear to be valid because its author thought it was (in which case, researchers are helping correct the misconception); does it appear to be valid but isn't because the author was sloppy (in which case, the author is wasting people's time — or the referees of the paper didn't reject it and should have); or perhaps the paper is in some sense fraudulent, and the author intended it to be published, sloppy or not.

Arguably, confusion between the different sorts of validity with respect to the status of a particular paper is *the* problem, certainly a bigger problem than errors, or even fraud *per se*. Confusion in the mind of a reader is worse than confusion (or ignorance) in the mind of a writer as there are usually many more readers than writers. For example, being knowingly inspired to do better is different (and far more constructive) than being misled. But this relies on correctly recognising the status of the paper. Even a fraudulent paper might inspire people. People *wanted* to research on cold fusion regardless of validity: they suspected the Fleischmann and Pons work [5] was fraudulent or exaggerated, but it gave the area a useful impetus and more funding *regardless*.

The difficulty of reproducing research will discourage researchers from trying to build on the foundations of published research; research methods will be understood less, they will be refined less (as fewer people try to use them), and new research will be isolated — and it will also be harder to assess.

In short, we should try to reduce errors, from whatever causes. However, as our knowledge is incomplete, some errors are inevitable: we should also try to improve the detectability of errors. Of course, our attitudes must be realistic and appropriate to purpose: in some areas, we want to be inspired, for instance by futuristic fiction which explores how things might be other than they are, but in

other areas, such as flight deck safety in aircraft, we want to be certain, so far as possible, to avoid errors and to make the detection of non-avoided errors as easy as possible. In science, also, we want to report results so that potential or actual errors in our work are as readily detectable as possible.

Notwithstanding Francis Bacon (truth will sooner come out from error than from confusion) [2] and others, Popper [13] was the first to present a systematic argument on the importance of being refutable — and of being *clearly* refutable by being sufficiently precise that errors could be spotted, rather than missed or dismissed. Gray and Salzman's classic though controversial paper [7,8] was an exposé of a widespread relaxed attitude to statistics and experimental method in human-computer interaction. A review of the *Journal of Machine Learning Research* suggests that about a third of its programs are not reproducible [17]; Mlodinow [10] recounts the Nobel Prize winner Richard Feynman's reaction to fraudulent physics, namely he was more concerned at its wasting the time of honest researchers — see also Feynman's discussion of radical honesty [4], which emphasises the central importance of doing science so that potential errors and assumptions are intentionally made clear rather than, as is common, concealed or ignored. There is a large literature on error in science, with [19] being a good review. In computing, tools are available to improve reproducibility [16], a paper that also includes substantial discussion of reproducibility in different disciplines. (I know of no such tools for HCI specifically.)

2 Handling an Error

David Mackay, Alan Blackwell and their colleagues have reported to me that there is an error in my own paper [15]. This particular error resulted from my sloppy proof reading of the paper, which is embarrassing, but I hope that is mitigated by the fact that the error could be, and indeed was, detected.

The Appendix of the present paper briefly summarises the error, and shows alternative approaches to how it can be corrected. Although the case is concrete and (at least to me) interesting, the details have been moved into a self-contained Appendix. The purpose of the present discussion is to generalise, rather than dwell on the specific issues of a particular paper, though of course that paper illustrates the principles.

In terms of the business of science, reporting and correcting a published error is no more than a footnote to a journal's wider business. On the other hand, the paper in question proposes not just a scientific idea advancing research in the field (e.g., under-pinning [18]), but the theory itself is an approach that can be developed further for practical user interface design. The detection and correction of an error in the paper is not just correcting a piece of science publishing, but can also be seen as a parable of detection and correction of errors in practical user interface design. Just as we do not want to mislead researchers, we do not want designers to use methods that allow them to be misled in real design projects: when researchers are misled, time is wasted; when designers are misled, bad systems will be built and lives will be risked. In other words, what at first sight is a criticism of the paper and its author (there was an error) in fact is an

argument providing support for applying the approach (the error *was* detected), certainly in safety related applications.

Detailed discussion of the error in the paper is provided in the Appendix A, and the discussion and lessons are summarised in Appendix B.

3 Different Sorts of Error

Although authors may take steps to disguise an error, or an error may be concealed or ignored by accident, in principle errors can be identified. We may distinguish between internal errors: errors that can be spotted by the internal evidence or arguments of paper; errors that can be spotted only by reference to external information (perhaps locked in lab books, or transient and lost); and errors of reportage, which can only be spotted, if at all, by reproducing experiments and collecting more data.

Quite different sorts of problem arise through vagueness and witholding information. Within these sorts of inadequacy, we can see variations:

- Inadequacy due to brevity. The paper is too short. The simplest solution here is to make good use of the internet or FTP to provide supplemental material.
- Inadequacy due to separation. The work was done too long ago (or the paper is being read by somebody some years after it was written). Details are now no longer available — particularly computer-based material. The solution here is to use media like digital libraries and journal repositories that may be archival, or at least far more likely to be archival than the author's resources permit of local storage.
- Due to sloppiness or disregard to standards, the work is vague.
- Due to exaggeration or 'clarification' the work as reported is in some ways better than was actually obtained.

4 Recommendations

This paper has reviewed the role of error in science publication (and has given a 'worked example' centred on and exploring the consequences of an error in one of the author's own HCI papers). So what?

Lessons can be drawn out of the discussion and example, which lead to recommendations for better practice.

4.1 Star Rating

First, it is important that there is a close match between the author's intentions and the reader's understanding of the status of the paper. As an extreme example: a paper written in jest is humourous if the reader recognises it as funny; and a serious paper would not be successful if the readers thought it a joke, and *vice versa* (notwithstanding [14])! A simple idea, then, is that papers should clearly indicate key features of their claim to validity. For example, a star rating could be used — as follows.

A paper that merely claims to be inspirational might have one star. The paper would be refereed on the basis of how inspiring it was, not how reliable it was. Maybe the ideas discussed do not quite work, but nevertheless they are very interesting. A two star paper claims, further, to have got something to work, but perhaps not everything. All the way to a five star paper that claims not only do the ideas work as described, but all background data and programs are available from a server. The exact definitions of the star ratings would depend on the journal (or conference) and the field. A mathematics paper, generally, makes an implicit claim to be five star in this sense — hence the error in my own paper was an issue, because it betrayed the implicit star rating.

Note that an author can improve the star rating of a paper. They can include more data or program code, or provide URLs for readers (and referees) to access the original information. There are many papers, in journals and conferences, that describe systems — but the systems are not available. One may wonder how the actual system implemented and the published paper conform. If we had a culture of awarding stars to papers, there would be a pressure to make papers and what they are about correspond more closely — and be more open to inspection. Indeed, the more stars, the more easily another researcher can build on or reproduce the original work.

Another way of viewing the star rating is answering the question, "does the system described work?" Almost everything in HCI is about an interactive system and the author's experience with it (or the author's experience of a user's experience with it), so something should have worked! So: zero stars for things that do not work; one star for something that almost worked, or worked well enough for a small experiment; two stars for something that really works — but has only be used for the purposes of the paper; three stars for something that not only worked for the paper, but has been developed to work elsewhere as well; four stars for something that has been rigorously tested elsewhere, on different platforms; and five stars for something that is supported to work well anywhere.

4.2 Triangulation

Secondly, authors (and editors and referees) should encourage triangulation: more than one way of justifying a result. If a paper is the only claim to the result, there is no triangulation. One takes the paper on faith (which may be expoited). Triangulation requires alternative routes to the same result — the simplest is that the paper provides URLs so that any reader of the paper can reconstruct for themselves the same results. The discussion of the matrix error above gave several ways in which the same result can be found.

In short, publishing and doing research in a way that promotes triangulation improves the assurance of the results, as well as giving the reader of the paper more choices in reproducing, understanding, or working from the claims made.

4.3 Data, Formal Argument, Programs, etc, Downloadable

Thirdly, many more formal papers in HCI (and papers aspiring to formality) present fragments of formal text. Often the fragments or the notations they are

written in are not fully defined. It is of course very hard to abstract out what needs saying for a paper; a full elaboration may take up excessive space. However, mostly, it is a reasonable expectation that the author has actually done the work that the paper abstracts. If so, the work should be available in full, for instance at a URL.

I know of no journal in HCI that has a working mechanism for reporting corrections to papers, let alone a means for encouraging the detection or correction of errors. (Conferences are inevitably in an even worse position.) Why don't journal web sites have explicit correction threads?[1]

As Altman [1] says, if journals are willing to publish subsidiary material on the web, they should explicitly tell authors. More so, journal articles would be improved if it was made clear to readers whether and to what extent the published paper is backed up by subsidiary material; this is a specific form of star rating.

Who would wish to publish papers that announce near their title banner that there is no supporting subsidiary material, if the paper clearly has the nature that there should have been such material (e.g., the paper discusses results obtained from a program; the program presumably exists and was at least once run)? No doubt authors would aspire to the greater prestige of having the right boxes ticked!

4.4 Further Work

Stylistically it is tempting to mix fact and vision. Often fiction is much clearer than the messy reality. What an author plans to do, planned to do, or would rather have done may make a lot more sense that what actually happened. Indeed it is sometimes recommended to write what you want to happen, so that expressing clear goals will guide experimental work; this obviously leaves open-ended the obligation to fix up the writing when the experimental work fails to deliver the original goals neatly.

In some fields, papers fit into standard patterns (e.g., "introduction; previous work; method; experiment; discussion; conclusion; references"). These standard patterns do not help factor out fact from wishes. Many papers, then, would be improved by having a section clearly labelled Further Work, or equivalent, so that the author can explain the simple vision without risk of misleading the reader.

4.5 Clarification and Communal Practice

Finally, we need to sort out these (or better) ideas, because many authors — and doctoral students — are working hard to advance our field, but they may fail in one of two ways:

- They may fail to publish because their notions of validity are not the disciplinary notions of their referees' or examiners'. We will call this the *cross-validity problem*.

[1] It was the lack of a working facility in the *ACM Transactions on Computer-Human Interaction* that stimulated the writing of this paper.

– In order to avoid the cross-validity problem (consciously or accidentally) authors may succeed in publishing invalid work that is hard to tell is invalid in any discipline.

4.6 Learning from Other Fields

HCI is not unique in its problems of validity; compared to medical fields, the debate surrounding Gray & Salzman [7] is tame! For example, von Elm and Egger lament the 'scandal' of epidemiological research [3]. Since problems in medical fields have had a longer history than in HCI, various standards have been developed such as the Consolidated Standards for Reporting Trials (CONSORT) and the Standards for the Reporting of Observational Studies (STROBE), etc. I believe it would be an advance if HCI developed or adopted such standards, so that they can be used where applicable — and so that authors can aspire to higher, and explicit, standards in the validity of their work.

Another suggestion from the medical field is post publication peer review [1]. Some HCI journals (such as *Interacting with Computers*), have had reviews, but these have not been sustained.

4.7 An Incomplete List ...

This list of recommendations is incomplete. It raises issues, and suggests solutions. There are many other issues, and other solutions. I hope the list stimulates the HCI community to address the problem of validity, whether incrementally or radically, whether starting from this list or by introducing new ideas. The benefits of improved validity are substantial, and the field clearly has the scope to improve.

5 Conclusions

Theories should be clear and robust enough that errors in their exposition (as in this case) or in their foundations can be reliably and robustly detected. The error reported and corrected in this present paper was essentially a typographical error rather than a conceptual error that needed correction for 'science to progress.' Instead, it can be used to make another point, about the practical application of theory. Had the error or a similar error been made in the design context, it could have been detected and rectified before a faulty product was put into production.

HCI is a very difficult and broad discipline. The difficulties we have in doing good work and reporting it accurately may lead to compromising validity — and to errors. By discussing errors and their role in publication, this paper also suggested some criteria for improving the detectability of errors, and improving the author: reader match of expectations of validity: requiring triangulation, and using a 'star rating' system. As well as a list of recommendations, which are of course of varying value in different subfields of HCI, we introduced the term *cross-validity problem* to enable the community to talk about the central issue explicitly.

To make any recommendations (such as the list above in this paper) work, ways must be found to make the recommendations *sustainable*. Currently, many economic and political factors conspire against improving validity. In the UK, the Research Assessment Exercise attaches no importance to reviewing work for maintaining or improving quality. Instead, it strongly emphasises the value of publishing, and therefore it must tend to increase the volume of publication, and, other things being equal, reduce the standards of validity in publication.

If we do not address validity (and the problem of cross-validity) in HCI we are losing sight of the point of HCI: to improve the quality of life of users, which will come about faster and more reliably through pursuing validity in the relatively abstract realm of research, publication and publication processes.

Acknowledgements

The author thanks David Mackay and Alan Blackwell, Cambridge University, for pointing out the error, the consideration of which triggered this paper. Jeremy Gow pointed out that MAUI was already powerful enough to detect the problem. Harold Thimbleby is a Royal Society-Wolfson Research Merit Award Holder, and gratefully acknowledges this support, which also supported the original research.

References

1. D. G. Altman, "Poor-quality medical research: What can journals do?" *Journal of the American Medical Association*, **287**(21):2765–2767, 2002.
2. F. Bacon, *Novum Oranum (The new organon or true directions concerning the interpretation of nature)*, 1620.
3. E. von Elm & M. Egger, "The scandal of poor epidemiological research," *British Medical Journal*, **329**:868–869, 2004.
4. R. P. Feynman, "Cargo Cult Science," in *Surely You're Joking Mr. Feynman!* R. Hutchings ed., Vintage, 1992.
5. M. Fleischmann & S. Pons, "Calorimetry of the Pd-D2O system: from simplicity via complications to simplicity," *Physics Letters A*, **176**:118–129, 1993.
6. G. Gay & H. Hembrooke, *Activity-centered design*, MIT Press, 2004.
7. W. D. Gray & M. C. Salzman, "Damaged merchandise? A review of experiments that compare usability evaluation methods," *Human-Computer Interaction*, **13**(3):203–261, 1998.
8. W. D. Gray & M. C. Salzman, "Repairing damaged merchandise: A rejoinder," *Human-Computer Interaction*, **13**(3):325–335, 1998.
9. J. Grudin, "Crossing the Divide," *ACM Transactions on Computer-Human Interaction*, **11**(1):1–25, 2004.
10. L. Mlodinow, *Some Time with Feynman*, Penguin Books, 2004.
11. J. Gow, H. Thimbleby & P. Cairns, "Misleading Behaviour in Interactive Systems," *Proceedings BCS HCI Conference*, **2**, edited by A. Dearden and L. Watts, Research Press International, pp33–36, 2004.
12. G. Lindgaard, "Is the notion of validity valid in HCI practice?" *Proceedings 7th International Conference on Work with Computing Systems*, pp94–98, 2004.

13. K. Popper, *The Logic of Scientific Discovery*, Routledge, 2002.
14. A. Sokal, "Transgressing the Boundaries: Toward a Transformative Hermeneutics of Quantum Gravity," *Social Text*, **46/47**:217-252, 1996.
15. H. Thimbleby, "User Interface Design with Matrix Algebra," *ACM Transactions on Computer-Human Interaction*, **11**(2):pp181–236, 2004.
16. H. Thimbleby, "Explaining Code for Publication," *Software — Practice & Experience*, **33**(10):975–1001, 2003.
17. H. Thimbleby, *Journal of Machine Learning Research*, *Times Higher Education Supplement*, 9 May, 2004.
18. H. Thimbleby, "Computer Algebra in User Interface Design Analysis," *Proceedings BCS HCI Conference*, **2**, edited by A. Dearden and L. Watts, Research Press International, pp121–124, 2004.
19. J. Waller, *Fabulous Science: Fact and Fiction in the History of Scientific Discovery*, Oxford University Press, 2004.
20. D. R. Wixon, "Evaluating usability methods: why the current literature fails the practitioner," *Interactions*, **10**(4):28-34, 2003.

A The Error

Ironically the error in question occurs in the discussion of a safety related interactive device [15, p217]. The user interface of a commercial Fluke digital multimeter is being discussed. The meter (like many user interfaces) has modes that change the meaning of buttons: in different modes, buttons mean different things. In particular the Fluke multimeter has transient modes entered by pressing shift keys: these change the device mode briefly, which is then restored after the next key press.

It suffices to quote an extract from the original paper as published, along with its original error:

The Fluke meter has a shift button, which changes the meaning of other buttons if they are pressed immediately next. (It only changes the meaning of three buttons, including itself, all of which anyway have extra meanings if held down continuously; additionally, the shift button has a different, non-shift, meaning at switch on.) In general if S represents a shift button and A any button, we want SA to be the button matrix we choose to represent whatever "shifted A" means, and this should depend only on A.

For any button A that is unaffected by the shift, of course we choose $SA = A$. Since the shift button doubles the number of states, we can define it in the usual way as a partitioned matrix acting on a state vector (**unshifted-state** ⋮ **shifted-state**). Since (at least on the Fluke) the shifted mode does not persist (it is not a lockable shift), all buttons now have partitioned matrices in the following simple form

$$\left(\begin{array}{c|c} A_{\mathsf{unshifted}} & \mathbf{0} \\ \hline \mathbf{0} & A_{\mathsf{shifted}} \end{array}\right)$$

and

$$S = \left(\begin{array}{c:c} \mathbf{0} & I \\ \hdashline I & \mathbf{0} \end{array}\right)$$

which (correctly, for the Fluke) implies pressing SHIFT twice leaves the meter unshifted (since the submatrices are all the same size and $SS = I$).

The error in the above description is that the matrix written as

$$\left(\begin{array}{c:c} A_{\text{unshifted}} & \mathbf{0} \\ \hdashline \mathbf{0} & A_{\text{shifted}} \end{array}\right)$$

should have been

$$\left(\begin{array}{c:c} A_{\text{unshifted}} & \mathbf{0} \\ \hdashline A_{\text{shifted}} & \mathbf{0} \end{array}\right)$$

This could be argued a trivial error for a scientific paper (a rate of 0.5% error reported per page), and one that is surrounded by context that makes the intention clear. However, had the same error been made in a real design, then the specified device would not behave as intended, perhaps catastrophically so.

That Mackay could spot the error is some encouragement that a designer, or a design team, could equally spot similar errors in an actual design process. How, then, might the error be detected — and are there more general lessons than the particular example?

For clarity, hereon we notate the correct matrix A and the erroneous matrix $\underline{A}$. The matrix, in either its correct or incorrect form, is clearly a composite of an unshifted and a shifted meaning. The differences between A and $\underline{A}$ appear in how the shifted meaning persists, or does not persist, as the button is pressed repeatedly by the user. $\underline{A}$ allows the shifted meaning to persist, which is incorrect.

We now present three very different ways of seeing this error. One is suitable for hand calculations; the next more suited to an automatic tool such as MAUI [11] (which can already detect this problem) or a computer algebra system [18]; finally, we show there is an informal approach to detect the error that would be open to any designer but (like all such approaches) suffers from the likelihood of false positive assessments.

A.1 A Straight Forward Calculation

The paper gives a recipe for constructing any matrix A from its shifted and unshifted meanings, A_{shifted} and $A_{\text{unshifted}}$. Since shift is not supposed to persist, for any two matrices A and B each constructed in the way suggested, the product AB should not mention B_{shifted}, since a shift before A could only affect A but not B.

If we follow the construction shown in the original paper, unfortunately B_{shifted} does appear in the product (it is highlighted by an arrow):

$$\underline{AB} = \left(\begin{array}{c|c} A_{\text{unshifted}} & \mathbf{0} \\ \hline \mathbf{0} & A_{\text{shifted}} \end{array}\right)\left(\begin{array}{c|c} B_{\text{unshifted}} & \mathbf{0} \\ \hline \mathbf{0} & B_{\text{shifted}} \end{array}\right)$$
$$= \left(\begin{array}{c|c} A_{\text{unshifted}}B_{\text{unshifted}} & \mathbf{0} \\ \hline \mathbf{0} & A_{\text{shifted}}B_{\text{shifted}} \end{array}\right)$$

Thus if A is shifted, B must be also, which is incorrect (though of course the whole matrix is wrong). Compare this result with the correct construction:

$$AB = \left(\begin{array}{c|c} A_{\text{unshifted}} & \mathbf{0} \\ \hline A_{\text{shifted}} & \mathbf{0} \end{array}\right)\left(\begin{array}{c|c} B_{\text{unshifted}} & \mathbf{0} \\ \hline B_{\text{shifted}} & \mathbf{0} \end{array}\right)$$
$$= \left(\begin{array}{c|c} A_{\text{unshifted}}B_{\text{unshifted}} & \mathbf{0} \\ \hline A_{\text{shifted}}B_{\text{unshifted}} & \mathbf{0} \end{array}\right)$$

Here, there is no B_{shifted} in the product anywhere; whether A is shifted or unshifted, the meaning of AB depends on $B_{\text{unshifted}}$ and not on B_{shifted} under any circumstances. This is what is meant by the shift not being persistent.

As an aside, it is interesting to note that we can examine the meaning of two consecutive key strokes without knowing what preceded them (or even the actual state of the device before they are pressed); indeed, in this case we know what AB means regardless of whether it follows S or not.

A.2 A Mode Based Calculation

The design tool MAUI [11], which Gow built for exploring properties of interactive systems specified by matrix algebra already has facilities for detecting this class of error. Here, we show how MAUI works.

It is important to remember that the mathematics is concealed by the tool. A designer using a suitable tool need not be as mathematically literate as the exposition here appears to suggest.

MAUI can find device properties automatically (such as the partial properties discussed above); relations between modes are just another case of the properties MAUI can handle. In particular, properties MAUI discovers about a device can be expressed in terms of modes and mode changes.

MAUI defines modes as sets of states. We would therefore define two modes, $\mathbf{s}$ and $\mathbf{u}$ representing the shifted and unshifted modes. The designer would be told that $\mathbf{u}\underline{A}$ remains in mode $\mathbf{u}$ but that $\mathbf{s}\underline{A}$ stays in $\mathbf{s}$. But $\mathbf{s}\underline{A}$ should have returned to mode $\mathbf{u}$!

Inside MAUI, this is how it is done: A mode is represented as a vector, such that for all states s in the mode M represented by $\mathbf{m}$, $\mathbf{m}_s = s \in M$. We define $\Box\mathbf{a} \sqsubseteq \mathbf{b} = \forall i{:}\, \mathbf{a}_i \Rightarrow \mathbf{b}_i$. It is now a routine calculation to show $\Box\mathbf{u}\underline{A} \sqsubseteq \mathbf{u}$ and $\Box\mathbf{s}\underline{A} \sqsubseteq \mathbf{s}$ (which is the error), whereas $\Box\mathbf{u}A \sqsubseteq \mathbf{u}$ and $\Box\mathbf{s}A \sqsubseteq \mathbf{u}$ (which is correct). Our notation is suggestive of counting states in a mode, and this is in fact what MAUI does.

A.3 Simulation

MAUI allows a device to be simulated (and many other tools can simulate devices specified by matrices or equivalent formalisms), and it is a simple matter for a designer to try out a simulated device out in order to satisfy themselves it behaves as intended.

The problem here is that any hand-driven simulation will likely miss errors — the designer might have been more worried over some other potential error, and omitted to test whether the shift key effect was persistent or not; or the designer might have found that the shift key works, but they have failed to check every possible combination of key presses. The state spaces of typical devices are enormous, and way beyond direct human assessment.

Though a simulation is realistic, a designer is really in no better a position than a user: just because the device appears correct in some or even in a majority of states, the designer is liable to believe the device correct on incomplete information. Worse, the areas of the device the designer explores carefully are likely to be areas of concern and hence are anyway the areas that have been more rigorously designed; problems may remain undiscovered in areas that no designer has paid much attention. For safety related devices, therefore, it is crucial that a tool-based or mathematical analysis is made. Indeed, if a designer 'plays' with a device simulation in MAUI and believes some property true, they can ask MAUI to confirm this or point out the conditions under which the property fails.

B Discussion

In short, the paper [15] claimed a property (shifted meanings do not persist) and showed a matrix that failed the claimed property, as is evident by the straight forward calculations carried out above. In the design context, perhaps the matrix A or $\underline{A}$ would be proposed, and would then be checked against the desired property or properties. Simply, $\underline{A}$ would fail the test and would be eliminated from the design.

Had a similar design issue (or claim in a scientific paper) been treated using, say, transition diagrams, which are a superficially simpler formalism, it is unlikely that the design property could have been checked, let alone analysed so readily. Matrices have the clear advantage of being easy to calculate with. Indeed the calculations needed above were routine and easy to do by hand (they only involved 2×2 matrices — regardless of the complexity or sizes of the submatrices A_{shifted} and $A_{\text{unshifted}}$).

Arguably the algebraic formula $\Box \mathbf{s}\underline{A} \sqsubseteq \mathbf{s}$ (or its straight forward translation into words: pressing the button a keeps the device in shifted mode) is a clearer representation of the error than the earlier result involving $\underline{AB}$, but the calculation using modes relies on being very clear about which states are in which modes, as well as doing a multiplication involving all states. Such calculations are better suited to a computer than a human!

In an ideal world, a real designer would probably use a design tool to handle or hide the details; understanding matrix multiplication and doing hand calcu-

lations would be largely and in some cases unnecessary. In a more reasonable, not so idealised world, the design task would probably be split between different people: the specification of design requirements (such as non-persistent shift meanings) would be formulated by mathematically competent designers once; then a design tool would be used to automatically check the required properties continued to hold as the design was developed or iterated — in this case, the development and continual modifications of the design could be managed by the tool without any reference to the underlying technical computations, matrix algebra or otherwise.