Preprint: final version available from

CONNELL, I., BLANDFORD, A. & GREEN, T. (2004) CASSM and cognitive walkthrough: usability issues with ticket vending machines. *Behaviour and Information Technology*. 23(5). 307-320.

CASSM and cognitive walkthrough: usability issues with ticket vending machines

Iain Connell
UCLIC
University College London
Remax House
31-32 Alfred Place
London WC1E 7DP
i.connell@cs.ucl.ac.uk
tel. +44 020 7679 5244
fax +44 020 7679 5295

Ann Blandford
UCLIC
University College London
Remax House
31-32 Alfred Place
London WC1E 7DP
a.blandford@cs.ucl.ac.uk

Thomas Green
greenery@ntlworld.com

We focus on the ability of two analytical usability evaluation methods (UEMs), namely CASSM (Concept-based Analysis for Surface and Structural Misfits) and Cognitive Walkthrough, to identify usability issues underlying the use made of two London Underground ticket vending machines. By setting both sets of issues against the observed interactions with the machines, we assess the similarities and differences between the issues depicted by the two methods. In so doing we de-emphasise the mainly quantitative approach which is typical of the comparative UEM literature. However, by accounting for the likely consequences of the issues in behavioural terms, we reduced the proportion of issues which were anticipated but not observed (the false positives), compared with that achieved by other UEM studies. We assess these results in terms of the limitations of problem count as a measure of UEM effectiveness. We also discuss the likely trade-offs between field studies and laboratory testing.

1. Introduction

This paper focuses on the ability of two analytical usability evaluation methods, namely CASSM (Concept-based Analysis for Surface and Structural Misfits) and Cognitive Walkthrough (Lewis et al. 1990, Polson et al. 1992, Lewis & Wharton 1997) to identify the usability issues with a walk-up-and-use ticketing system. We characterise the differences between the two methods in terms of their ability to depict the range and type of issues which were taken to underlie a large number of observed sub-optimal system-user interactions with the ticketing system.

CASSM -- previously known as Ontological Sketch Modelling (Blandford & Green 1997, Blandford & Green 2001, Connell et al. 2003) -- will be described below. Cognitive Walkthrough is an analytical usability inspection method which focuses on the ability of a system or device to support user goals at different stages of each interaction task. The ticketing system is the two main London Underground ticket vending machines.

A potential source of difficulty experienced by the user in a user-system interaction is any mismatch, or misfit, between the user's conceptualisation of the system or device and that which the system or device imposes on the user. We believe that the identification of user-system misfits is an important goal for any analytical approach to usability evaluation. Misfit analysis arises naturally out of the CASSM approach via that method's focus on the differences between device and user models. Cognitive Walkthrough (CW) was chosen for comparison with CASSM because of its depiction of the support for user goals. In that method, any gulfs between designers' and users' views of the support given by the system for user actions at each stage of a task, and between system and user goal structures, can be considered as misfits. In that sense, the two approaches can be considered as complementary. However, CASSM is different from CW in moving the focus away from tasks and actions towards system and user concepts.

Most studies which have made comparisons between usability evaluation methods (UEMs) have relied on usability problem count as the dependent variable. We shift the emphasis away from problem count towards a more direct comparison of the types of usability issues and user-system misfits which the two methods identify. By setting both sets of misfits against the observed interactions with these machines, we assess the degree to which the issues depicted by CASSM were substantially the same as or different from those depicted by CW. In focusing more on the issues themselves than their numerical differences, we de-emphasise the mainly quantitative approach which is typical of the comparative UEM literature. We also assess the results in the light of Gray and Salzman's (1998) critique of this literature, which identified several shortcomings in the problem-count approach to UEM comparisons.

1.1 CASSM and misfit analysis

Central to CASSM's approach is the simultaneous representation of the user's conceptualisation of the domain and that which is built into the system or device. In a CASSM analysis the analyst describes both the concepts (entities and their attributes) pertaining to the device or system, and those relating to the user's model of domain use. Description is also made of the concepts built into the user-system interface. Concepts which are present in the system but absent from the user model, or present in the user model but absent from the system, are a clear source of system-user misfits. Concepts which are present in either the system or user models but only apprehended with difficulty at the interface will be a cause of further misfits whose severity depends on the level of that difficulty.

CASSM makes the distinction between surface and structural misfits. Surface misfits are those relating solely to features of the interface between system and user, whereas structural misfits concern changes made to the underlying structure of a system or data model. An example of a surface misfit is the navigational difficulties engendered by lack of feedback or non-reversibility of user actions (so that users lose their place in a series of state changes and cannot easily get back to where they were). An example of a structural misfit is the difficulty in engendering an accurate user model of style hierarchies in a word processor (so that the effects of resetting style attributes may be difficult to predict). It is the nature of ticketing systems that the majority of misfits to be identified will be surface rather than structural.

A result of a CASSM analysis is a table of each concept's user-interface-system dependencies. This table can be used to reason about the type and likelihood of misfits relating to each concept, in terms of whether that concept is present (P), absent (A) or difficult (D). A taxonomy of P, A and D combinations can be used to characterise the type and severity of each misfit, with reference to the specifics of the device and system in question. A description of the CASSM taxonomy can be found in Blandford & Connell (2004a).

CASSM may be applied to a large range of systems and applications, from walk-up-and-use installations like vending machines to work-systems such as the London Ambulance Service dispatch system (Blandford et al. 2002).

1.2 Cognitive Walkthrough and Misfit Highlighting

Though Cognitive Walkthrough (CW) does not deal explicitly in misfits, it does focus on support for user goals at different stages of user-system tasks. In order that each stage may be supported by the system or device, there must be an implicit match between what actions the device requires of the user and the user's expectations for how to proceed towards the current goal. If there is no such match, either the system or the user must bridge the gulf (in Norman's (1986) terms, of execution and evaluation) by some means: in the system case, supplying informative feedback on more profitable strategies; in the user case, bringing to bear knowledge or insights from similar instances in different domains. Thus CW can be considered as a tool for highlighting potential user-system misfits at every task stage. However, its lack of encouragement of exploration beyond the tasks under consideration, and its relatively low-level attention to the details of user action and system response (Sears 1997), may leave it wanting in comparison with the wider scope which we hope that CASSM encourages.

UEM studies which have involved cognitive walkthrough include Desurvire et al. (1992), Cuomo and Bowen (1994) and Sears (1997). These and most other UEM studies have made use of relatively low numbers of users (in total, or per group). The low user populations, combined with the difficulty disentangling different causal factors out of whole-UEM comparisons, make it difficult to extrapolate from these particular studies to more general statements about CW (or the other UEMs against which it was compared). Thus we cannot easily tell from such studies whether CW is consistently better than heuristic evaluation (Nielsen 1994) in predicting real usability issues, or vice versa; or, if it is (better than heuristic evaluation), precisely what it is about CW that makes it so. As Gray & Salzman (1998) have argued, one approach (to these issues of both external and internal validity) would be to limit the claims made for UEM studies and to be more precise about which method variants are used. Gray & Salzman also criticised the use of problem count in UEM studies. The direction of this paper will be away from problem count towards a more direct comparison of the types of usability issues which can be uncovered by the two methods under investigation.

1.3 The ticket machine observations

The currently reported observational study took place in 2002. Observations of ticket machine use were made in similar fashion to those in two previous studies by the first author (1990-91 and 1996, reported in Connell (1998)), namely overall usage tallies followed by detailed machine-user interactions (see Section 3). However, whereas the previous studies were concerned with error analyses, the focus of the current study moves away from errors towards the wider context of the London Underground ticketing system. In this context, 'errors' can be viewed as alternatives to the optimal task sequences which were envisaged by the machines' designers. In that sense, the many sub-optimal interaction sequences are viewed more in the light of system-user mismatches than as discrepant or erroneous user actions.

By setting the observed interactions against the kinds of mismatches, or misfits, which CASSM and CW would predict for these particular machines in this wider context, we assess the potential ability of the two methods to support reasoning about such misfits. Thus we explore the ways in which two approaches which claim to say something about user and designer goals, system and user concepts, have the potential to be descriptive of real user behaviour.

In the next Section we describe the two vending machines. Section 3 describes the results of the observations and analyses, and Section 4 the matching process through which analyses and observations were compared. In Section 5 we summarise and discuss the results of the comparisons.

2. The London Underground ticket vending machines

All London Underground stations have at least one of each of two main types of ticket vending machine (hereafter referred to as TVMs), namely the FFM or Few Fare Machine (figure 1) and the MFM or Multi Fare Machine (figure 3). Typically there are two FFMs to each MFM, many outer zone stations having a single MFM. Since the introduction of the machines in the late 1980s, the FFM has changed very little. The MFM was little modified before 1996, but considerably changed by 2001, when all of its hard buttons were replaced with a touch-screen monitor.

2.1 FFM (Few Fare Machine)

The FFM is a quick-use coins-only machine for regular users who know the price and type of their ticket. Its ten hard price buttons (return tickets on the leftmost buttons, single tickets the rightmost) offer only the most popular tickets from the station in question, including a one-day Travelcard (combined underground, train, bus and tram ticket). On this machine the first two of the three user steps can be performed in either order. Only one ticket at a time can be purchased on the FFM. Only one child ticket is commonly available.

[Insert figure 1 about here]

Figure 2 shows typical ticket buttons and labels on the FFM in 2002.

[Insert figure 2 about here]

2.2 MFM (Multi Fare Machine)

The MFM offers tickets to all underground, Tramlink and DLR (Docklands Light Railway) stations (and some overground stations). The post-2001 MFM features a single touch-screen monitor in place of the previous hard button arrays of ticket types and destination stations, plus a credit/debit card slot in addition to the note tray and coin slot. In this paper we focus on the layout of the touch screen interface as it was in 2001-2002 (the 2003 version includes icons showing the current status of change availability, card & note acceptance, and a new form of destination selection). In general, the 2001 and later MFM interface is like its predecessors in requiring selection of ticket type and destination before payment can be made. The MFM offers a range of ticket types including one-day Travelcards as well as Singles and Returns. Child tickets, multiple tickets and receipts are also available.

[Insert figure 3 about here]

Figures 4 and 5 show screens 1 & 2, and typical screens 3 & 4, of the 2002 MFM touch screen interface.

Screen 1 (figure 4) features an initial ‘attention grabber’ display, and screen 2 contains the top-level menu. The ‘Change Given’ (etc.) panel in all four screens shows the current change, notes and cards acceptance status.

Screen 3 (figure 5) shows the result of selecting the ‘Single to Zone 1’ button in screen 2. Screen 4 shows the result of then selecting the ‘Press here for multiple tickets’ button in screen 3, and then the Adults: ‘3’ button in screen 4 itself. The ‘You have selected:’ panel confirms the current selection.

[Insert figure 4 about here]

[Insert figure 5 about here]

3. Observations and evaluations

3.1 Observations of user-machine interactions

Observations took place over two to three weeks in September of 2002, at three stations (Bounds Green, Arnos Grove and Highbury & Islington). These were performed by the first author. Sessions took the following form:

1. Overall tallies of machine vs. ticket window use, successes and failures. A failure was deemed to be an interaction sequence which did not involve ticket retrieval. These included cancellation with return of money inserted, and abandonment of a machine in mid-sequence. A total of 560 machine attempts (211 FFM, 349 MFM) were observed. Of these, 7.6% of FFM and 11.7% of MFM attempts were recorded as failures. (It was not possible to also record detailed usability issues at this stage.)

2. Further detailed observations of machine use, including tallies of usability issues, whether critical or non-critical, plus user comments and notes. A critical issue was deemed to be one whose outcome was a failure, typically that which immediately preceded cessation of the sequence. A non-critical issue was one which did not lead to failure, typically involving recovery and continuation of the sequence. Thus the same issue might be deemed to be critical in one case and non-critical in another. A total of 165 sub-optimal usability incidents (35 FFM, 130 MFM) were recorded during this stage. Of these, 25 FFM (71%) and 58 MFM (45%) incidents, the latter including six whose causes could not be attributed, were noted as

being critical. (Unattributable MFM issues may have included experimentation with the interface, and price-comparison.)

Table 1 shows the complete set of observed issues, grouped into critical and non-critical and sorted by combined (FFM plus MFM) incidence. It can be seen that the most frequent critical issues were concerned with the availability of destinations or ticket types, plus not noticing the ‘closed’ sign (both machines are regularly closed for short periods for re-stocking with change). The most frequent non-critical issues concerned inefficient and incorrect selection sequences. Lowest frequencies were single occurrences.

[Insert table 1 about here]

3.2 CASSM analysis

Two independent CASSM analyses were carried out, by the first and second authors. The resulting pair of analyses were combined by agreement between the analysts into a single set of issues, presented in table 2. As stated above, CASSM allows assessment to be made of the seriousness of each misfit, by means of a table of user-interface-system dependencies whose cell contents are judged to be present (P), Absent (A) or Difficult (D). For example, the misfit ‘Traveller age limits (Adult, Child) are absent from interface ... ‘ was judged to be serious by both analysts, while ‘Payment acceptance status ... not clear when at machine‘ was judged to be minor by analyst 1 but serious by analyst 2. Issues are listed in order of seriousness, as judged by one or both analysts. The corresponding tables of dependencies can be found in Blandford & Connell (2004a).

For the purposes of matching the CASSM and CW issues against the observed issues, assessment has been made of the effect(s) of each issue, as manifested in likely user actions with one or both machines. These appear in the ‘May lead to’ column of table 2. In order to ease comparison with the items in table 1, these descriptions have been couched, where possible, in behavioural terms. For example, ‘Traveller age limits (Adult, Child) are absent from interface and may not be known to user’ was deemed to lead specifically to ‘Purchase of Child or Adult ticket where not necessary or not permitted’, whereas the attributed outcomes of ‘Available ticket types are difficult to determine’ were more open-ended (on the

MFM), leading to 'More flexible or cheaper ticket type not found', and/or 'More time taken to select a ticket type'. The latter is an example of the sometimes one-to-many relationship between attributed causes and observable outcomes, analogous with Hollnagel's (1993) distinction between genotypes and phenotypes.

It is acknowledged that some of the behavioural consequences thus attributed were not directly observable, either because no record was made of such an event, or because they rely on knowledge of the user's intentions (and no such knowledge had been elicited). For example, 'Current time not displayed in interface ...' may have led to 'Purchase of higher price ticket than necessary (by waiting until off-peak time', but without further clues (in this case, to the effect that the user either had not known about off-peak times, did not know when they came into effect, or had wrongly thought that the threshold time had passed), it was not possible to confirm this outcome by mere observation. By contrast, in 'Traveller age limits (Adult, Child) are absent from interface...', it may have been possible to determine that a Child or Adult ticket had been bought in error (as opposed to by design), since Child ticket purchases are clearly indicated in the interface (by a change of background colour from blue to yellow). The issue of descriptive level is important for this paper and will be taken up in the Discussion.

[Insert table 2 about here]

3.3 Cognitive walkthrough analysis

A Cognitive Walkthrough (CW) analysis was carried out by the second author, using the illustrations and MFM sample screens shown in figures 1 to 5 plus background information (e.g. on ticket type validity) provided by the first author. With the agreement of analyst 2, the first author extracted a set of issues from the text description of the CW; these are presented in table 3. As with the CASSM results above, attempt has been made to account for the effects of each issue in behavioural terms, and to indicate where these were not directly observable. Once more, some effects were specific to each machine. The text of the CW analysis can be found in Blandford & Connell (2004b).

[Insert table 3 about here]

4. Matching analyses to observations

The accuracy of each analytical method was assessed by matching the issues listed in tables 2 and 3 (Sections 3.2 and 3.3) against the observations in table 1 (Section 3.1). This was done by agreement between the two analysts, particularly in relation to differences in interpretation concerning the two sets of CASSM issues. The result of these comparisons are presented in tables 4a to 4c. Table 4a shows the issues which were shared by both methods. Tables 4b and 4c respectively are those anticipated by CASSM and CW alone.

4.1 Hits, possible hits, and false positives

[Insert table 4a about here]

[Insert table 4b about here]

[Insert table 4c about here]

Only definite and unambiguous matches were judged to be Hits (H). Less clear-cut and ambiguous (but nevertheless plausible) matches were labelled Possible Hits (PH). Issues which were anticipated by either CASSM or CW but for which no match could be found were the False Positives (FP); those which had been judged to have Not Directly Observable consequences were marked NDO. Issues which had been observed but which neither CASSM or CW had anticipated -- the Misses -- will be presented in the next Section. Once again, some issues were judged to have more than one potential match.

We can see that while there was some agreement between CASSM and CW on the salience of the payment acceptance status (the dot matrix sign was above rather than facing the user while she or he was at the machine: see figures 1 and 3) and the available destinations, a third (concerning the absence of destination zones) was judged to be only a possible hit. A hit concerning the place of destination went to CASSM, but CW had four further hits which CASSM did not. CW had more false positives than CASSM, while CASSM produced more issues than CW which could not be directly observed. The numbers of possible hits were comparable between the two methods.

On this analysis, then, CW was better than CASSM in unambiguously identifying observed usability issues, whereas CASSM had anticipated fewer issues than CW which had not manifested during the observations. Both methods were about the same in flagging up potential or possible issues whose precise manifestation had not (or not yet) been spotted, or those which could not be confirmed without further elicitation of user intentions and/or comments.

4.2 Misses

The hits, false positives and NDOs together accounted for around two thirds of the observed issues. The remainder, those issues which neither method had anticipated yet which had been observed, were the Misses. These are listed in table 5.

[Insert table 5 about here]

In this case distinction has been made between critical and non-critical issues, as defined in Section 3.1. As there explained, an issue might be critical in one instance and non-critical in another, so that two issues (‘Coins or notes rejected ...’ and ‘Wait by Machine showing ...’) feature in both categories.

It can be seen that most of the missed issues were critical, two concerning mechanical faults which were not noticed or not apparent. The two most frequent non-critical issues also involved mechanical or, broadly, hardware configuration problem which had therefore not been picked up by either CASSM or CW. Nor did either method account for the continuing incidences of a phenomenon first noticed in Connell (1998), namely that users would abandon one machine (predominantly the MFM) for the other, having selected a ticket and ascertained the price, even when the MFM is giving change (and accepting coins of the same denomination as the FFM). We shall merely remark on the persistence of this possibly unique feature, while reserving comment on the other issues for the Discussion.

5. Discussion

This attempt to match CASSM and CW misfits against observed usability issues has, on the face of it, not shown any strong advantage of one method over the other. Neither method was particularly good at identifying misfits which unambiguously manifested in observed user-system breakdowns (critical issues) or non-critical usability issues. In terms of definite matches (hits), CASSM depicted fewer usability issues than CW but had fewer false positives, while the two methods shared a small number of hits. However, both methods produced a relatively strong emphasis on issues that could not be directly observed, and there were a disappointing number of merely potential matches between the analytic and empirical processes (in Gray & Salzman's (1998) terms). In addition, neither method directed attention to the types of mechanical problems which are typical with vending machines. The latter is a failing which we believe they would share with other analytical evaluation methods which focus on internal (human and software) processes rather than external (hardware and I/O) ones.

However, the decision to direct attention at the likely consequences of the anticipated issues (the 'May lead to' columns of tables 2 and 3), and thus pitch both sets of descriptions at the same or comparable level, did bring about a low number of false positives, when contrasted with other studies which have compared predicted and observed usability problems. In such studies, 'usability inspection' and 'user testing' have been considered to address problems of different scope (Bailey et al. 1992, Desurvire 1994, Karat 1994, Karat 1997), and false positive estimates have ranged from as little as 5% to 49% and even 82% of predicted problems (Cuomo and Bowen 1994, John and Marks 1997). In the current study, the tactic of delineating cause-effect relationships for analytical issues appears to have succeeded in reducing the FPs to a single incidence (CASSM) and a mere four out of 18 issues (CW). Compared with the 20% reduction in FPs reported by Cockton et al. (2003) over a previous study by the same group (Cockton and Woolrych 2001), to an apparently <u>improved</u> figure of 50%, the approach appears to pay dividends. This seems to be so even though Cockton et al. (2003)'s problem report format specifically took account of consequences and contexts (as well as causes) of usability problems. As for the NDOs, we do not have a figure against which to compare these estimates, but in the current study the FPs and NDOs together accounted for just 41% and 42% of all issues anticipated by each method.

Why, then, have false positives been so distressingly high, for those who have measured them ? What is it about the usability inspection process that leads analysts to over-estimate the difficulties and problems surrounding a user-system interaction process, compared with what (as far as we can tell) actually happens ? Our view is that false positive are a feature of a more general phenomenon of asking people what they think, that is: they will tell you -- and experts may have a tendency to tell you everything that could go wrong, rather than all that might, merely, under typical conditions. It is not so much that experts tell more than they know (Nisbett and Wilson 1977), they just know more than there is. Put another way, in order to catch each single instance of every possible problem which an expert throws up we may have to wait for much longer than two weeks, if we spot them at all.

On this interpretation, false positives and NDOs would appear to be endemic in accounts of interaction processes produced by expert analysts, unless encouraged to prioritise on grounds such as severity (S) or frequency of occurrence (F_o). One tactic is to incorporate both factors into an overall priority estimate (P) by using their product: $P = S \times F_o$. Unfortunately, however, F_o is rarely available (if we knew what it was we wouldn't need to ask experts to guess at it), so the temptation is to resort to the safer but much inferior frequency of prediction (F_p), so that $P = S \times F_p$ for issues predicted by more than one analyst. But the evaluator effect (Hertzum and Jacobsen 2001) attests to the very lack of agreement between multiple evaluators on what are even the most serious issues, so that F_p is the least reliable factor in such circumstances. It appears, then, that though experts may well be best placed to distinguish between 'real' and 'merely possible' problems, we (meaning comparative UEM researchers) have omitted to ask them.

As noted in Connell (1998), the level at which usability problem descriptions are grouped can make a considerable difference when comparing one set of problems (or issues) against another. In that study, the tactic of grouping problems at a higher level achieved an apparent increase in the predictive power of an error analysis method called DEA (Christie et al. 1995) from 50% to 85%. In the current study, pitching descriptions at a lower rather than higher level appears to have had little effect on predictive power but a substantial effect on extraneous predictions. The possibility therefore arises of a trade-off between higher level but less precise descriptions (increases hits but may not reduce FPs) and lower level descriptions with consequence considerations (little effect on hits but reduces FPs). This is interestingly similar, in part at any rate, to Cockton et al. (2003)'s initial conjecture that raising the number of analysts (and thus the description level ?) will increase hit rate at the expense of producing more FPs.

The extent to which the manipulation of the results of the two analytical processes was a device engineered in (in order to bring down the number of FPs, or merely replace FPs with NDOs) is for future work to address. Methods such as CASSM and CW can, by their nature, lend only general insights into a particular issue or potential misfit which may (or may not) manifest in a range of sub-optimal user-system interactions. In contrast, a behavioural or observational account can only deal in specific sub-optimal instances. Thus it is hardly surprising that the necessary generalities of a CASSM and CW do not match unambiguously onto the specifics of an observational account without some tweaking. The high proportions of potential hits and NDOs compared with unambiguous hits is a testament to such a process.

The above argues for a better means of interrogating user-system interactions than is offered by mere observation. In a user testing (laboratory) scenario, the usability engineer has the opportunity to query test users on their actions, perhaps involving retrospective playbacks of the interaction sequence. In field studies, there is little opportunity to do so, and, ethical considerations aside, attempting to query real users about their use of ticket vending machines, in real time (when the main intent is to board a train, not talk to people with clipboards) is fraught with difficulty. (This study did involve some attempts at interviewing, with as little success as those in Connell (1998).) It is ironic that the very advantage of a field study, with all the additional context and situational background that it affords, may be the main source of disadvantage in disambiguating user behaviour. While the compromise solution of a laboratory simulation or machine setup (with volunteer test users in semi-realistic situations) can go some way to addressing these difficulties, it is unlikely that the broader contextual insights that methods like CASSM and CW can bring to a real-world study would be noticed.

For example, we have seen that both CASSM and CW might direct attention at missing destination zone information on both FFM and MFM. This issue would be unlikely to be picked up in user testing which is structured around tasks such as ‘Buy a return ticket to zone 1’ or ‘Find the price of a single to Acton Town’. However, such tests <u>would</u> allow detection of wrong selections (e.g. Acton Central rather than Acton Town) and machine malfunctions (e.g. incorrectly printed tickets), both legitimate errors which would not be caught by non-interventional field observations. Nor would purely observational field studies identify those users who choose the ticket window over the TVMs because they know (or merely suspect) that the machines do not enable them to make sufficiently informed choices. Thus while user testing might be best at error-detection and post-test interrogation of user actions (both of which miss out on the contextual richness of CASSM or CW), and field observations capture a large range of real user behaviours not amenable to testing, neither the analytic nor empirical approaches employed in this paper are sufficient, alone or together, to encompass the whole picture. And, as we have seen, there is no necessary and straightforward relationship between the two.

What, then, might CASSM and CW really bring to an understanding of London Underground ticket vending user-system misfits ? For example,

'Destination zone may not be known by user (and zone maps on machines are small and often worn)' (table 2)

and

'A choice of Adult or Child tickets is enforced on users early in the interaction (combined Adult/Child tickets can be bought later)' (table 3)

tell us that (a) an essential part of the very information that the user needs in order to make optimal use of both machines, and achieve the maximum flexibility in later travel decisions, is missing from the machines themselves, and (b) parents with children, who need the maximum of assistance in making multiple purchase decisions, are forced into a premature commitment -- making them act as child purchasers in order to buy adult and child tickets together -- which they will probably revoke later. But neither of these figure in the observed interactions, precisely as described, because, quite properly (for reasons of good empirical practice), observations were made without prior recourse to the analytical accounts. Instead, we have shoehorned the potentially rich CASSM and CW descriptions into crude cause and effect relationships, in order to make a point.

In this paper we have begun to account for some of the reasons why a comparison of usability evaluation methods which relies on problem counts for its main measure of method effectiveness may be found wanting. We concur with Gray and Salzman's (1998) assessment of the UEM literature, in that closer attention needs to be given to what insights the methods under study are capable of bringing to the evaluation process, rather (we would say) than focusing on how much of it they do in numerical terms. In our view, the quantitative approach is flawed in its attempt to apply statistical comparisons to ordinal data whose unit of measurement -- the usability problem -- is widely divergent in character both within and between applications. We have also used the attempt to demonstrate some qualitative differences between the two methods to make some proposals regarding the level at which usability issue descriptions, and their consequences, are pitched, and the likely trade-offs between real-world field studies and the more contrived scenarios of laboratory testing.

Acknowledgements

The assistance and co-operation of London Underground Ltd. ticketing and marketing staff, particularly Robin Indar-Singh of the former, and staff at Arnos Grove, Bounds Green and Highbury & Islington stations is gratefully acknowledged.

The work reported in this paper has been supported by EPSRC grant no. GR/R39108.

References

BAILEY, R.W., ALLAN, R.W. and RAIELLO, P., 1992, Usability testing vs. heuristic evaluation: a head-to-head comparison. In *Proceedings of the Human Factors Society 36th Annual Meeting*, 1992, Volume 1 (Human Factors Society), pp. 409-413.

BLANDFORD, A.E. and CONNELL, I.W. (2004a). CASSM Taxonomy of User-Interface-System Dependencies plus Example Analyses. Available (posted March 2004) at: http://www.cs.ucl.ac.uk/staff/i.connell/DocsPDF/taxonomy.pdf

BLANDFORD, A.E. and CONNELL, I.W. (2004b). Cognitive Walkthrough of London Underground Ticket Vending Machines. Available (posted March 2004) at: http://www.cs.ucl.ac.uk/staff/i.connell/DocsPDF/cw.pdf

BLANDFORD, A.E. and GREEN, T.R.G., 1997, OSM: An ontology-based approach to usability evaluation. In *Proceedings of an International Workshop on Representations in Software Development*, Queen Mary and Westfield College, July 1997, pp. 82-91.

BLANDFORD, A.E. and GREEN, T.R.G., 2001, From tasks to conceptual structures: misfit analysis. In *People and Computers XV - Interaction Without Frontiers (Volume II)*, joint proceedings of HCI 2001 and IHM 2001, Lille, September 2001, edited by J. Vanderdonckt, A. Blandford and P. Gray (Toulouse: Cepadues-Editions), pp. 113-116.

BLANDFORD, A.E., WONG, B.L.W., CONNELL, I.W. and GREEN, T.R.G., 2002, Multiple viewpoints on computer supported team work: a case study on ambulance dispatch. In *People and Computers XVI*, proceedings of HCI 2002, London, September 2002, edited by X. Faulkner, J. Finlay and F.D. Étienne (Springer-Verlag), pp. 139-156.

CHRISTIE, B., SCANE, R. and COLLYER, J., 1995, Evaluation of human-computer interaction at the user interface to advanced IT systems. In *Evaluation of Human Work: a Practical Ergonomics Methodology*, 2nd edition, edited by J.R. Wilson and E.N. Corlett (London: Taylor & Francis), pp. 310-356.

COCKTON, G. and WOOLRYCH, A., 2001, Understanding inspection methods: lessons from an assessment of heuristic evaluation. In *People and Computers XV - Interaction Without Frontiers*, joint proceedings of HCI 2001 and IHM 2001, Lille, September 2001, edited by J. Vanderdonckt, A. Blandford and P. Gray (Toulouse: Cepadues-Editions), pp. 171-191.

COCKTON, G., WOOLRYCH, A., HALL, L. and HINDMARCH, M., 2003, Changing analysts' tunes: the surprising impact of a new instrument for usability inspection method assessment. In *People and Computers XVII - Designing for Society*, proceedings of HCI 2003, Bath, September 2003, edited by E. O'Neill, P. Palanque and P. Johnson (London: Springer), pp. 145-161.

CONNELL, I.W., 1998, Error analysis of ticket vending machines: comparing analytic and empirical data. *Ergonomics, 1998, **41(7)**, 927-961*.

CONNELL, I.W., GREEN, T.R.G and BLANDFORD, A.E., 2003, Ontological Sketch Models: highlighting user-system misfits. In *People and Computers XVII - Designing for Society*, proceedings of HCI 2003, Bath, September 2003, edited by E. O'Neill, P. Palanque and P. Johnson (London: Springer), pp. 163-178.

CUOMO, D.L. and BOWEN, C.D., 1994, Understanding usability issues addressed by three user-system interface evaluation techniques. *Interacting with Computers, 1994, **6(1)**, 86-108*.

DESURVIRE, H.W., 1994, Faster, cheaper !! Are usability inspection methods as effective as empirical testing ? In *Usability Inspection Methods*, edited by J. Nielsen and R.L. Mack (New York: John Wiley & Sons), pp. 173-201.

DESURVIRE, H.W., KONDZIELA, J.M. and ATWOOD, M.E. (1992), What is gained and lost when using evaluation methods other than empirical testing. In *People and Computers VII*, proceedings of HCI '92, York, September 1992, British Computer Society Conference Series 5, edited by A. Monk, D. Diaper and M.D. Harrison (Cambridge: Cambridge University Press), pp. 173-201.

GRAY, W.D. and SALZMAN, M.C., 1998, Damaged merchandise ? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 1998, **13(3)**, 203-261*.

HERTZUM, M. and JACOBSEN, N.E., 2001, The evaluator effect: a chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction, 2001, **13(4)**, 421-443*.

HOLLNAGEL, E., 1993, The phenotype of erroneous actions. *International Journal of Man-Machine Studies, 1993, **39**, 1-32*.

JOHN, B.E. and MARKS, S.J., 1997, Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology, 1997, **16(4/5)**, 188-202*.

KARAT, C-M., 1994, A comparison of user interface evaluation methods. In *Usability Inspection Methods*, edited by J. Nielsen and R.L. Mack (New York: John Wiley & Sons), pp. 203-233.

KARAT, J., 1997, User-centered software evaluation methodologies. In *Handbook of Human-Computer Interaction*, 2nd edition, edited by M. Helander, T.K. Landauer and P. Prabhu, (Amsterdam: Elsevier Science B.V.), pp. 689-704.

LEWIS, C. and WHARTON, C., 1997, Cognitive walkthroughs. In *Handbook of Human-Computer Interaction*, 2nd edition, edited by M. Helander, T.K. Landauer and P. Prabhu, (Amsterdam: Elsevier Science B.V.), pp. 717-732.

LEWIS, C., POLSON, P., WHARTON, C. and RIEMAN, J., 1990, Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Empowering People: CHI '90 Conference Proceedings*, ACM conference on human factors in computing systems, Seattle, April 1990 (Special issue of the SIGCHI Bulletin), edited by J.C. Chew and J. Whiteside (Seattle: Addison-Wesley), pp. 235-242.

NIELSEN, J., 1994, Heuristic evaluation. In *Usability Inspection Methods*, edited by J. Nielsen and R.L. Mack (New York: John Wiley & Sons), pp. 25-62.

NISBETT, R.E. and WILSON, T.D., 1977, Telling more than we can know: verbal reports on mental processes. *Psychological Review, May 1977, **84(3)**, 231-259*.

NORMAN, D.A., 1986, Cognitive engineering. In *User Centred System Design*, edited by D.A. Norman and S.W. Draper (Hillsdale, NJ: Lawrence Erlbaum Associates), pp. 31-62.

POLSON, P.G., LEWIS, C., RIEMAN, J. and WHARTON, C., 1992, Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies, 1992, **36(5)**, 741-773*.

SEARS, A., 1997, Heuristic walkthroughs: finding the problems without the noise. *International Journal of Human-Computer Interaction, 1997, **9(3)**, 213-234*.

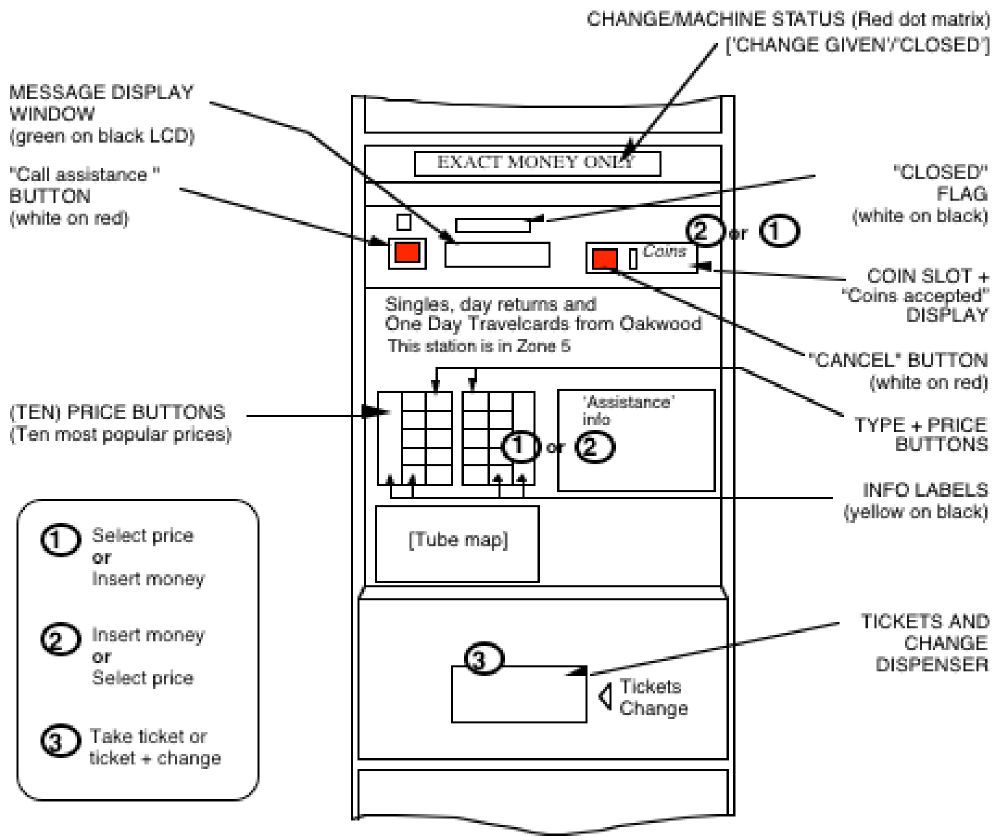


Figure 1: FFM (Few Fare Machine). Details as of October 2001, Oakwood station. Not to scale.

(Information labels)	(Price labels)	(The buttons)
Adult single to any station in Zone 4 (not via Zone 1) and selected local stations	Adult Single £1.00	(button)
Adult single to any station in Zone 3 (not via Zone 1)	Adult Single £1.30	(button)
Adult single to any station in Zone 2 (not via Zone 1)	Adult Single £1.70	(button)
Adult single to any station in Zone 1	Adult Single £2.70	(button)
Child single to any station in Zone 4 (not via Zone 1)	Child Single 40p	(button)

(The buttons)	(Price labels)	(Information labels)
(button)	Adult Return £2.00	Adult day return to any station in Zone 4 (not via Zone 1) and selected local stations
(button)	Adult Return £2.60	Adult day return to any station in Zone 3 (not via Zone 1)
(button)	Adult Return £3.40	Adult day return to any station in Zone 2 (not via Zone 1)
(button)	Adult Return £5.40	Adult day return to any station in Zone 1
(button)	Adult Day Travelcard (off-peak) Zones 1-4	Adult One Day Travelcard. Unlimited Tube, Bus, DLR, Tramlink and rail travel in Zones 1-4

Figure 2: FFM information labels, price labels and ticket buttons. Details as of September 2002, Arnos Grove station.

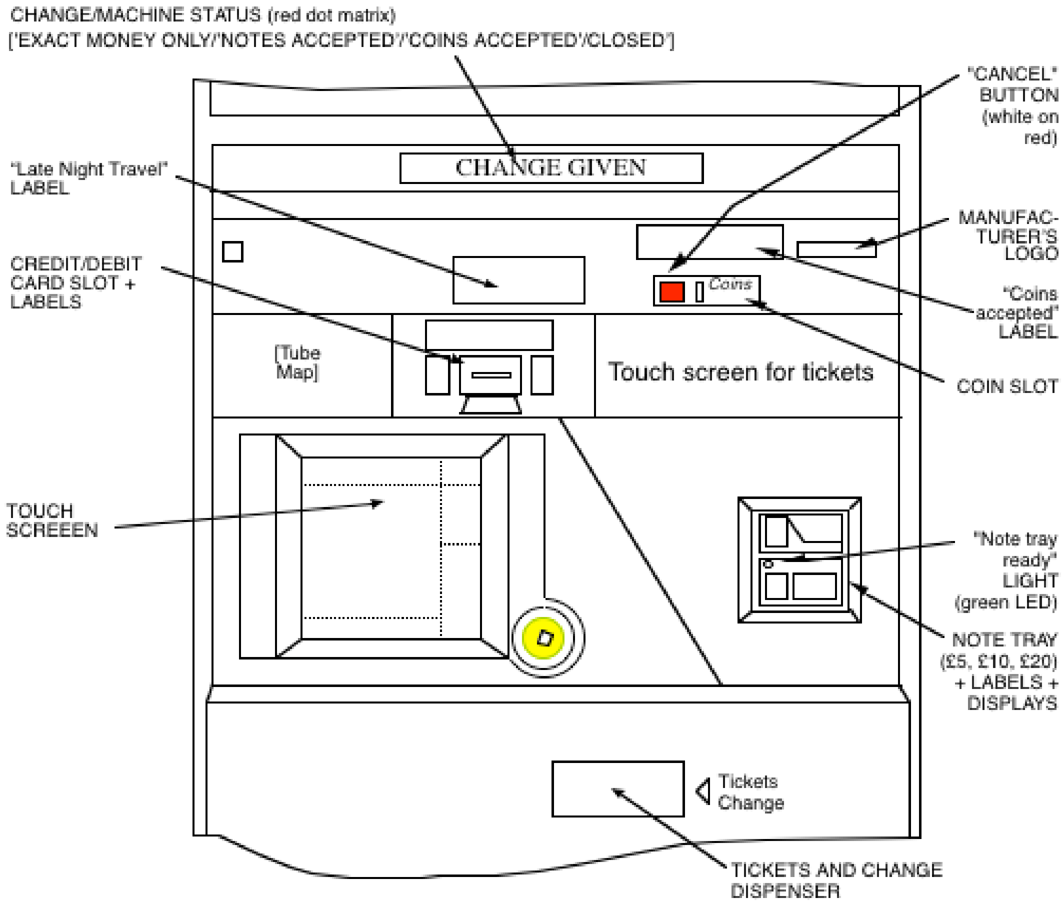
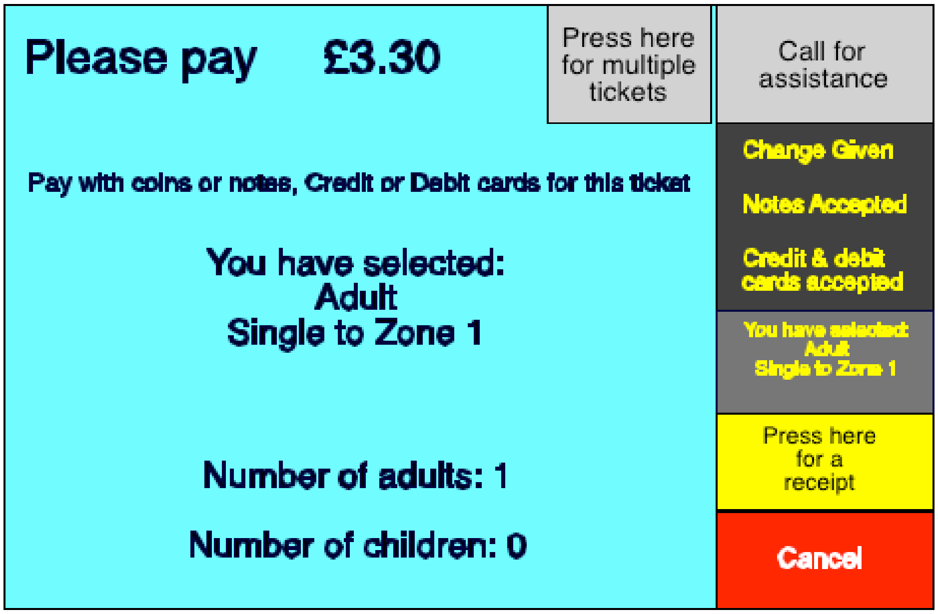


Figure 3: MFM (Multi-Fare Machine). Details as of October 2001, Oakwood station. Not to scale.

Figure 4: MFM 2002, screens 1 and 2. Details as of September 2002, Oakwood Station.

['Single to Zone 1' selected]



['Press here for multiple tickets', then Adults: '3' selected]

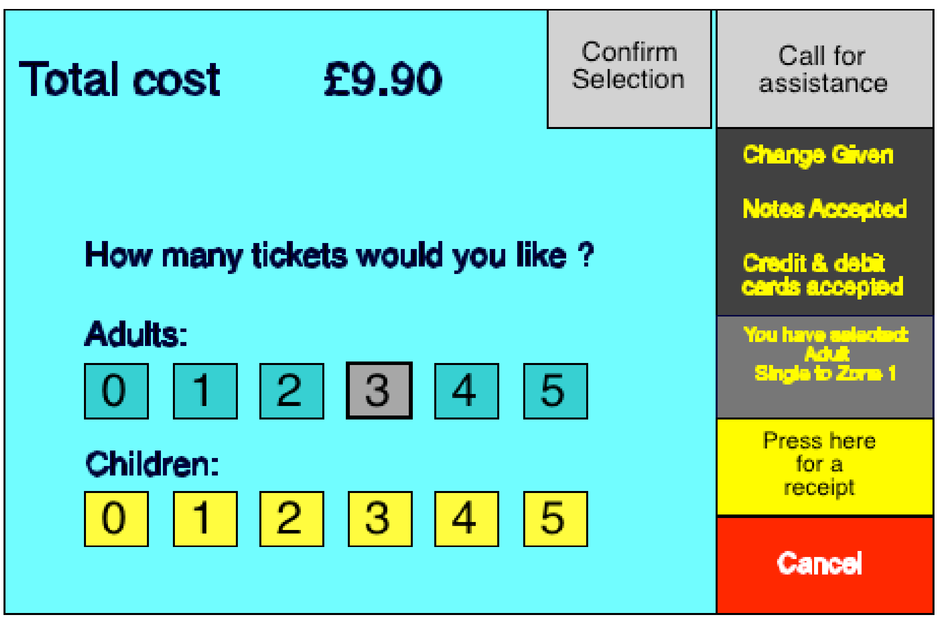


Figure 5: MFM 2002, screens 3 and 4. Details as of September 2002, Oakwood Station.

Observed critical issues (Analyst 1)	FFM	MFM
Desired destination/zone/Travelcard type not available		12
Closed or Out of Service status not noticed	9	2
Desired ticket type not available	10	
Payment (cards, notes, coins) acceptance status not noticed	1	8
Abandon one machine for the other even though the first is giving change	1	8
Coins or notes rejected (mechanical fault)	2	6
User does not have sufficient change of correct type	2	4
Unknown critical		6
Wait by Machine showing (not Cancelled from previous user)		3
Repeated transactions for multiple tickets (card refused on repeat attempt)		2
Timeout after delay in user actions		2
£20/£10 note not used (when change given)		2
Abandon one machine for the other when the first is not giving change		2
Wrong selection(s) (ticket type, station, zone, etc.)		1
Observed non-critical issues (Analyst 1)	**FFM**	**MFM**
Inefficient selection route (ticket type, zone, etc.)		19
Wrong selection(s) (ticket type, station, zone, etc.)	2	10
Coins or notes rejected (mechanical fault)		6
Repeated transactions for multiple tickets		6
Attempt to insert note/coin into card slot		6
Attempt to insert coin(s) or card of type not accepted		5
Unknown non-critical		4
Timeout after delay in user actions	1	3
Forget to press Child ticket button at start		4
Repeated selections, superfluous re-start	2	1
Payment (cards, notes, coins) acceptance status not noticed		2
User does not have sufficient change of correct type	2	
Wait by Machine showing (not Cancelled from previous user)	1	1
Card inserted wrong way round		2
Pressed (yellow) smart card button (not in use at time of study)		2
Desired destination/zone/Travelcard type not available (select nearest equivalent)		1
Coin slot jammed	1	
Call Assistance pressed for Cancel	1	

Table 1: Observed critical and non-critical issues, both TVMs, all stations, September 2002 (Analyst 1). Issues are sorted by incidence (FFM and MFM combined), within group.

CASSM issues	Analyst 1	Analyst 2	May lead to
Traveller age limits (Adult, Child) are absent from interface and may not be known to user	Serious (D-A-P)	Serious (D-A-A)	Purchase of Child or Adult ticket where not necessary or not permitted
Current time not displayed in interface (At peak time, TravelCards are more expensive than off-peak)	Serious (P-A-P)	Serious (D-A-P)	Purchase of higher price ticket than necessary (by waiting until off-peak time) (Not directly observable)
Day of travel not specified in interface (Cannot be other than current day, but this may extend past midnight)	Serious (P-A-P)	Serious (D-A-P)	Attempt to purchase a ticket for a day other than the current day (Not directly observable, but likely to result in failure)
Permitted travel method with particular ticket types (e.g. TravelCard, Single, Return) are not shown on interface	Serious (P-A-P)	Serious (A-A-P)	Purchase of ticket type which does not cover intended travel method (Not directly observable)
Source zone is not specifically shown on interface and may not be known to user	Serious (D-A-P)	Serious (A-P-P)	FFM: failed attempts to identify correct ticket/route (via Zone 1 or not via Zone 1)
Payment acceptance status (Cards, Notes, out of change) not clear when at machine	Minor (P-D-P)	Serious (A-P-P)	Attempt to insert card, note or coins when not currently accepted or machine is out of change
Destination zone may not be known by user (and zone maps on machines are small and often worn)	Minor (D-P-P)	Serious (A-P-P)	Failed or repeated attempts to select correct zone-combination as destination
Destination station difficult to select (cf. destination zone) because hidden on MFM interface	Minor (P-D-P)	Serious (P-A-P)	MFM: More time taken than necessary in specifying a destination (cf. zone or zones)
Place of actual destination (served by which destination station(s)) absent from both interface and system		Serious (P-A-A)	Purchase of ticket to wrong destination (Not directly observable)
			MFM: failed attempts to locate unavailable destination station
No indication in interface or system of actual distances between stations or places		Serious (P-A-A)	Purchase of expensive ticket for short distance journey (Not directly observable)
Peak/off-peak times not specified in interface. (User must intuit whether higher peak time prices are in force)		Serious (A-D-P)	Purchase of higher price ticket than necessary (Not directly observable)
Source station absent from interface	Serious (P-A-P)		FFM: Failed attempts to identify correct ticket/route (via Zone 1 or not via Zone 1)
Underground line(s) used for journey are not present in system or interface (and maps on machines are small)	Serious (D-A-A)		MFM: more time spent than necessary in locating destination zone (on separate map)
			FFM: extra attempts to identify correct ticket/route (via Zone 1 or not via Zone 1)
Available ticket types are difficult to determine	Minor (D-D-P)	Minor (P-D-P)	MFM: More flexible or cheaper ticket type not found
			MFM: More time taken to select a ticket type
User may not know the price of the ticket in advance (and may only later discover that she/he does not have sufficient change)	Minor (D-D-P)		Failed attempt at cash payment with insufficient coins or coins plus note(s)
The range of destinations (ticket types on FFM) available are not well displayed	Minor (P-D-P)		Attempt to choose a destination which is not available on that particular machine

Table 2: CASSM (analysts 1 and 2): issues and their likely consequences for 2002 TVMs, both machines. User-interface-system dependencies are P: Present, A: Absent, D: Difficult. Issues are ordered by combined (analyst 1 and 2) and sole seriousness assessment.

CW issues (Analyst 2)	**May lead to**
Payment acceptance status (Cards, Notes, out of change) not clear when at machine	Attempt to insert card, note or coins when not currently accepted or machine is out of change
	Abandoned transaction
The range of destinations available from each machine are not well displayed	Attempt to choose a destination which is not available on that particular machine
Information about the destination station zone is absent from both machine interfaces	Failed or repeated attempts to select correct zone-combination as destination
Both machines enforce a choice of zone plus Travelcard choice early in the interaction	Abandoned attempts at destination station selection
	Purchase of ticket to wrong zone (Not directly observable)
Both machines may fail to prevent the user from forgetting to take her/his change at the end of the transaction	Change left in machine
The functions of the different payment input devices (card, notes, coins) are not well differentiated	FFM: forestalled attempt at card or note transaction
	MFM: use of single payment method when a combination of methods is possible
	MFM: inappropriate use of notes or card slot
Machine has insufficient change, even though not yet showing Exact Money Only	All money inserted returned, no ticket given
FFM: Many reasonable goals (e.g. multiple tickets, tickets to specific local destination stations) are not supported or explicitly represented	Abandoned attempt to achieve a goal
	Goal incorrectly formulated in order to make it fit what is possible (Not directly observable)
FFM: Selected ticket is not confirmed in interface (and ticket delivered immediately when sufficient payment made)	Unwanted ticket purchased (Not directly observable)
	Unwanted ticket selected at start: transaction cancelled and re-started
FFM: Not clear that can pay first (rather than ticket selection first)	Predominant ticket-selection first
FFM: (Hard) buttons are slightly separated from price and information labels	Attempt to press label instead of button
FFM: When paying first, there is no indication of what ticket(s) are available for that amount	Repeat attempt using ticket-selection first
MFM: Available ticket types are difficult to determine	More time than necessary to select a ticket type
	More flexible or cheaper ticket type not found (Not directly observable)
MFM: Peak/off-peak times are not made explicit on the machine interface	Purchase of higher price ticket than necessary (Not directly observable)
MFM: A choice of Adult or Child tickets is enforced on users early in the interaction (combined Adult/Child tickets can be bought later)	Repeated purchases of Adult then Child tickets (or vice versa) rather than a single purchase of multiple (adults plus children) ticket
MFM: Multiple Tickets button is relatively low-profile on screen layout (and presented at same time as Please Pay instruction)	Repeated transaction(s) for additional tickets
	Failure of next transaction, if the first was by card (only one per card per day permitted)
MFM: Confirm Selection button on multiple ticket purchase is relatively low-profile	Cancelled or repeated transaction (for multiple tickets)

Table 3: Cognitive Walkthrough (analyst 2): issues and likely consequences for 2002 TVMs.

CASSM and CW Issue	**Match**	**Observed Issue**
Payment acceptance status not clear when at machine	H	Payment acceptance status not noticed
		Go to other machine
The range of destinations (ticket types on FFM) available are not well displayed	H	Desired destination /zone/Travelcard type not available
		Desired ticket type not available
Destination zone may not be known by user and is absent from both machines	PH	Inefficient selection route
		Wrong selection(s)

Table 4a: CASSM and CW: shared hits (definite matches) and possible hits resulting from comparisons of CASSM and CW issues with the observed issues. H = Hit, PH = Possible Hit.

CASSM Issue	Match	Observed Issue
Place of actual destination absent from interface and system	H	Desired destination/zone/Travelcard type not available
Available ticket types are difficult to determine	PH	Inefficient selection route
		Wrong selection(s)
Destination station difficult to select because hidden on interface	PH	Inefficient selection route
		Wrong selection(s)
Source zone is not shown on interface and may not be known to user	PH	Wrong selection(s)
Source station absent from interface	PH	
Underground line(s) used for journey are not present in system or interface	PH	Wrong selection(s)
		Repeated selections, superfluous re-start
User may not know the price of the ticket in advance	PH	User does not have sufficient change of correct type
Traveller age limits (Adult, Child) are absent from interface	FP	No match with observed issue
Current time not displayed in interface	NDO	Not directly observable
Day of travel not specified in interface	NDO	
Permitted travel method with particular ticket types are not shown on interface	NDO	
Place of actual destination absent from both interface and system	NDO	
No indication of actual distances between stations or places	NDO	
Peak/off-peak times not specified in interface	NDO	

Table 4b: CASSM: hits (definite matches), possible hits and false positives resulting from comparison of CASSM issues with the observed issues. H = Hit, PH = Possible Hit, FP = False Positive, NDO = Not Directly Observable.

CW Issue	Match	Observed Issue
A choice of Adult or Child tickets is enforced on users early in the interaction	H	Repeated transactions for multiple tickets
Multiple Tickets button is relatively low-profile on screen layout	H	
Confirm Selection button on multiple ticket purchase is relatively low-profile	H	
The different payment methods are not well differentiated on each machine	H	Attempt to insert note/coin into card slot
Both machines enforce an early choice of zone plus Travelcard	PH	Desired destination/zone/Travelcard type not available
Many reasonable goals are not supported or explicitly represented	PH	Repeated selections, superfluous re-start
		Timeout after delay in user actions
Selected ticket is not confirmed in interface	PH	Wrong selection(s)
Available ticket types are difficult to determine	PH	Inefficient selection route
		Wrong selection(s)
Multiple Tickets button is relatively low-profile on screen layout	PH	Repeated selections, superfluous re-start
The different payment methods are not well differentiated on each machine	PH	User does not have sufficient change of correct type
When paying first, no indication of what ticket(s) are available for that amount	PH	Repeated selections, superfluous re-start
Machines may fail to prevent the user from forgetting to take her/his change	FP	No match with observed issue
Machine has insufficient change though not showing Exact Money Only	FP	
Not clear that can pay first (rather than ticket selection first)	FP	
(Hard) buttons are slightly separated from price and information labels	FP	
Both machines enforce an early choice of zone plus Travelcard choice	NDO	Not directly observable
Many reasonable goals not supported or explicitly represented	NDO	
Selected ticket is not confirmed in interface	NDO	
Peak/off-peak times are not made explicit on the machine interface	NDO	

Table 4c: CW: hits (definite matches), possible hits and false positives resulting from comparison of the CW issues with the observed issues. H = Hit, PH = Possible Hit, FP = False Positive, NDO = Not Directly Observable.

Issues	FFM	MFM
	Critical	
Closed or Out of Service status not noticed	9	2
Abandon one machine for the other even though the first is giving change	1	8
Coins or notes rejected (mechanical fault)	2	6
Wait by Machine showing (not Cancelled from previous user)		3
	Non-critical	
Coins or notes rejected (mechanical fault)		6
Attempt to insert coin(s) or card of denomination or type not accepted		5
Forget to press Child ticket button at start (done later in sequence)		4
Wait by Machine showing (not Cancelled from previous user)	1	1
Pressed (yellow) smart card button		2
Coin slot jammed	1	
Call Assistance pressed for Cancel	1	

Table 5: Observed issues which were missed by both CASSM and CW analyses.