

Helping Viewers Press the Right Buttons: Generating Intuitive Labels for Digital Terrestrial TV remote controls

Jane Lessiter*, Jonathan Freeman, Robert Davis and Andrew Dumbreck*

i2 media research limited
Department of Psychology
Goldsmiths College
University of London

*External R&D Contract Manager
Technology, R&D
Ofcom
London

ABSTRACT

This paper describes a four phase research study into the ease of use of DTT (digital terrestrial television) remote controls, focusing on button labelling. The aim was to identify intuitive remote control button labels, thereby enhancing the ease of use of DTT remote controls. In phase 1, a series of brainstorming (discussion) groups was run involving members of the British public to generate intuitive button labels for a variety of DTT functions. The second phase involved expert screening of these labels. The revised lists of labels for the functions tested were then compiled into a questionnaire and distributed to members of the general public in phase 3. Respondents were asked to rank order their favourite labels for each function. The final phase involved exploring the extent to which the subjective preferences derived from the results of phase 3 mapped onto viewers' behavioural responses. Key measures explored included the speed and accuracy of correct button label identification. Generally, the results revealed close correspondence between the subjective preference data and behavioural responses. Whilst these results suggest that subjective preferences for labels are reasonable indicators of behavioural proficiency, we recommend that behavioural measures also be collected where possible. The approach reported here with reference to remote control button labels can be applied to a wide range of interactive media products and services

Keywords: *Digital Terrestrial Television, button labeling, remote control.*

Received 24 October 2003, received in revised form 9 December 2003; accepted 20 December 2003.

1. Introduction

Evidence from previous research and anecdotal reports suggest that people are often confused by the button labels used to convey particular functionality on remote

* Corresponding Author:
Jane Lessiter,
i2 media research limited,
Department of Psychology,
Goldsmiths College,
University of London,
New Cross,
London SE14 6NW, UK
E-mail: J.Freeman@gold.ac.uk/J.Lessiter@gold.ac.uk

controls for a range of interactive media products and services. The issue is particularly important with reference to digital TV, given the ubiquity of television and the benefits that easy to use digital TV can provide for large sections of the population, both in terms of increased choice of programming, and access to interactive entertainment and information services. Whilst digital TV penetration in the UK is high relative to other countries, for people who have not yet adopted digital TV to readily adopt it, it needs to be perceived as having benefits to the user and as being easy to use. DTV affords increased functionality over the offerings of conventional analogue TV. Potentially it could provide added benefits to users for whom TV is already a valued product in the home, with increased entertainment in the form of, for instance, more choice, more specialist offerings, and more time shifting ability to suit the individual's lifestyle. DTV can also provide a useful information resource, giving access to government and local service information via a familiar and trusted medium. However, the added functionality of enhanced interactive programme features can also increase the complexity of the system. In this context usability is of paramount importance, especially so for a product like television which has almost universal penetration and is generally perceived as being very easy to use. Earlier work funded by the UK's Independent Television Commission (Freeman & Lessiter, 2001) explored the perceived ease of use of dTV and interactive TV (iTV) relative to 17 other products pre-rated (n = 51) for ease of use. The results revealed that the ease of use of dTV and iTV were perceived to be on a par with personal computers – and fell among the bottom third *least* easy to use of the rated products. In contrast, analogue TV was rated amongst the top third most *easy* to use products.

The ease of use of digital TV, and perceptions relating to its ease of use, are of particular importance when it is considered that the sensory, cognitive and dextrous abilities of the general population - and their confidence with technology - vary substantially. Further, because of the typical uses of TV - entertainment, relaxation and information - and the context in which it is typically used (the home environment), people do not expect to expend effort in learning how to use it. If dTV can satisfy the expectations viewers have of television, based on their experiences with analogue TV, and it is perceived as being easy to use, viewers' positive experiences will generate positive word of mouth for digital TV generally and improve non-users' perceptions of digital TV. Improved ease of use will also enable viewers to fully exploit the

information and entertainment capabilities of digital and interactive TV, making its use a satisfying, enjoyable and fun experience.

In 2002, a large scale survey was conducted which aimed to identify what characterised people with the biggest ease of use problems (Freeman, Lessiter, Williams & Harrison, 2003). The survey, based on over 1300 people, asked respondents about their attitudes to technology, media consumption and dTV. Following factor and cluster analyses of the resultant attitudinal data, seven clusters of people were identified representing different types of media consumer. The clusters differed in their perceptions of the ease of use of technology generally and dTV in particular, the hurdles they faced in dTV adoption and their motivations to adopt digital TV. Whilst the segmentation was based purely on attitudinal measures, the clusters identified were strongly differentiated in terms of dTV ownership. Across the clusters, dTV penetration ranged from 5.9% ('Strong Resisters') to 65% ('Technology Enthusiasts'). The seven clusters were further categorised into three higher order groups differentiated primarily by their perceptions of the ease of use of digital TV. The three groups were labelled 'It's too slow and clunky for me', 'I can deal with it if there's content I want', and 'It's too complicated for me'.

A series of usability trials were then run (in collaboration with the Consumers' Association) followed by focus group discussions about participants' experiences with the trial dTV services (Freeman, Lessiter, Williams & Harrison, 2003). These trials were conducted to explore the relationship between the 'type' of user and the extent to which they were affected by dTV usability issues. Very generally, while 'It's too complicated for me' users were more likely to make internal attributions about the usability issues, 'It's too slow and clunky for me' users were more likely to blame the inadequacies of the system. However, many specific usability issues were identified as problematic, either for themselves and/or for others, by all users irrespective of grouping. Thus, priority areas on which to concentrate further research resource were identified. Digital TV remote controls, which have been referred to as a viewer's handshake with the product (Daly-Jones, personal communication), were identified as a key priority. For instance, some DTT users did not understand the button labels on many remote controls in that the labels were sometimes deemed unintuitive or even misleading as shown by the following quote:

“Also the exit button, I mean usually when you’ve finished something and you want it to stay on something you’re used to pushing ‘enter’ or ‘save’ or whatever, to push ‘exit’ just seems very bizarre in my mind...”

This, combined with industry requests to assist in their work to generate recommended labelling for a variety of key dTV functions, prompted our research into identifying what labels (which could be words or symbols) are most simply and easily understood by the majority of the public to convey different functions that digital TV offers (i.e., what labels are *intuitive*?). The use of intuitive labels on remote controls promises to reduce manufacturers’ and retailers’ product support costs. For example, telephone help-line use and product returns should fall if products are designed to be more intuitive and easier to use.

2. Method

To meet our research goal of identifying intuitive remote control button labels a four phase research programme was conducted. This involved both qualitative and quantitative methods, and subjective and objective measures.

2.1 Phase 1: Brainstorming Sessions

First, a series of five brainstorming sessions - each involving (five to six) members of the British public - was run to generate ideas for labels for nine different dTT functions (‘Subtitles’, ‘Languages’, ‘Text’, ‘Exit Text’, ‘TV guide’, ‘Return to Sound and Vision [from any interactive application]’, ‘Back to Previous Screen’, ‘System Set-up Menu’, and ‘Power On’).

Fourteen males and thirteen females (mean age = 51.7 years, SD = 23.4) participated and were recruited from (a) our database of ‘willing participants’, (b) students and staff from Goldsmiths College, and (c) several elderly social groups in the south-east London area. Seven participants reported having dTV (3 Freeview, 4 Sky Digital). The groups used people with a range of attitudes towards technology – from the very confident (‘It’s too slow and clunky for me’) to the technophobic (‘It’s too complicated for me’). There was a slight deliberate bias towards those with average (‘I can deal with it if there’s content I want’) and low confidence (‘It’s too complicated for me’). It was reasoned that people with greater confidence in using technology would be more likely to explore and eventually learn the button labels regardless of any

ambiguities. Group members were simply asked to think of any labels (symbols or words) relating to the given function description.

The number of suggestions for each of the dTV functions ranged from 6 ('Text') to 33 ('Return to Sound and Vision [from any interactive application]')

2.2 Phase 2: Expert Screening of Generated Labels

In phase 2 an expert screening panel comprising six members of the UK Government's Digital Television Project Technology and Equipment Group – Converter subgroup (TEG-C) inspected the list and deleted unsuitable suggestions. Suggestions that were too long for a button, ambiguous (or bizarre!), and that might be confused with labels for other functions were removed. The panel proposed minor modifications to some of the retained labels – such as presenting word labels in lower case, removing hyphens between words, and converting symbols to international standard versions. They also requested a few additions to the list to be tested in phase 3.

2.3 Phase 3: Questionnaire Survey – Ranking Button Labels

Phase 3 involved a questionnaire survey where the screened list of labels (from Phase 2) were presented for each function and respondents were required to rank order their favourites (the number of choices ranged from 4 [for 'Text'] to 12 [for 'Return to Sound and Vision']). The 'Subtitles-Languages' and 'Power On' functions were removed from this survey (low priority), but the 'TV Guide' function was divided into two functions – 'Now and Next' and 'Electronic Programme Guide' (EPG). The same list of labels that had been generated and screened for 'TV Guide' were presented for each. There were eight functions in total that were addressed in the questionnaire – 'Subtitles', 'Text', 'Exit Text', 'TV Guide' (i.e., Electronic Programme Guide), 'Now and Next', 'Return to Sound and Vision', 'Back to Previous Screen', and 'System Menu'. Two versions of the questionnaire were constructed – the labels were presented in a different order in each to reduce the possibility of the presentation order of labels affecting respondents' ranking decisions.

Five hundred and fifty questionnaires were distributed (a) across England, Wales and Scotland via post to people who had previously taken part in our other research (but not Phase 1 of this research project) and who were willing to take part in further

research, and (b) to miscellaneous people around Goldsmiths College. As an incentive, participants were offered entry into a prize draw to win £100. Of the 550 distributed questionnaires, 237 were returned and were suitable for data analysis (43% response rate). Responses were received from people residing in 23 of the 39 counties in England. In addition, four people from Scotland and one from Wales responded. The age of the sample ranged between 17 and 92 years (mean age = 59 years, SD = 19). Forty-four percent of the sample were male. Thirty-six percent of the sample reported having dTV at home, a figure slightly lower than that for the UK population as a whole.

Each nominated label was given a score corresponding to its ranking (e.g., the label ranked first favourite was given a score of 1). Unranked labels were treated as missing data. As not all respondents gave all five ranks (or for the 'Text' function, 4) subjective preference was calculated by a simple tally of respondents' first choices for each label. Thus, only the results for first favourite were analysed.

Overall the results suggested that text labels were preferred to symbolic labels. For no functions did a symbol receive the highest number of first favourite nominations. Clearly it is an empirical question as to whether this subjective preference translates into behaviour. Interestingly, using a small sample of visually impaired users, the Royal National Institute for the Blind (Gill & Perera, 2003) found that visually impaired people responded significantly faster to symbolic rather than text labels.

For three functions ('Text', 'Back to Previous Screen' and 'Subtitles') there were clear 'winners' where the label directly corresponded to its function (labels 'text', 'back' and 'subtitles'). Each of these labels received over 50% of first favourite nominations and indicated that people tend to prefer labels that do exactly what they say, rather than abbreviations. It may simply be a matter of finding the shortest best descriptor word for that function.

For three of the remaining functions, the differences between first and second choices were marginal with around a 5% difference in frequency of first favourite nominations. For the 'Return to Sound and Vision' function, the label 'return' was closely followed by 'TV' (30.1% vs. 26%), for 'TV Guide-Electronic Programme Guide' function the label 'guide' preceded 'progs' (35.4% vs. 29.2%), and for 'System Setup

Menu' function the label 'setup' marginally received more nominations than 'settings' (29.9% vs. 27.6%). For these functions, which may be less well conceptualised by users, it may be more difficult to identify unique descriptor terms.

2.4 Phase 4: Experimental Evaluation of Button Labels

Aims

The aim of the final phase was to explore the extent to which the subjective preferences revealed in phase 3 translated into behavioural performance – i.e., are the most favoured labels more quickly and accurately identified than those less frequently nominated? It aimed to explore intuitiveness (on first exposure), memorability (learnability) and common mistakes. In addition, this phase measures the degree of consistency between the results found in phase 3 and the behavioural paradigm being used here. Careful design of this phase was crucial, and in order to reliably achieve its aims with a finite resource it was necessary to minimise the number of functions and corresponding labels to be tested.

Selection of Functions and Labels to be Tested

Three functions were selected for testing in this phase – 'Subtitles', 'Return to Sound and Vision' and 'Back to Previous Screen'. These functions were chosen for various reasons. 'Subtitles' was selected because a dedicated button for this function is to be recommended by the TEG-C group. While the label in full can fit on some buttons, an alternative shorter label might be preferred by some manufacturers, and further, the range of label possibilities for this function also included a symbol. 'Back to Previous Screen' is a function for which there has been much debate as to the most suitable label. Finally, the 'Return to Sound and Vision' function was selected because there has been discussion in the industry regarding the introduction of this new 'comfort' button, and its inclusion in TEG-C's baseline receiver specification for Digital Terrestrial Television converter boxes has been discussed. As a new and potentially frustration-reducing function it seemed important to test labels for this unfamiliar function to explore whether participants would understand the label's meaning.

Three labels for each of the three selected functions were chosen based on the results from phase 3. Labels were tested that were both close contenders for favourite (i.e., where there were two very 'popular' labels), and also subjectively disparately favoured labels (i.e., the most and least frequently ranked favourite labels). For example, for the 'Subtitles' function, the clear winner – 'subtitles' (ranked favourite by

61% of the sample) - was compared with 'subT' (10%) and the internationally recognised 'subtitles' symbol (9%). For the 'Return to Sound and Vision' function, the labels 'return' (31%), 'TV' (26%) and 'out' (0.5%) were tested. And for the 'Back to Previous Screen' function, the labels 'back' (52%), 'back up' (2%) and 'cancel' (1%) were compared.

Design

A 3 by 3 mixed measures design was used with one within group factor (function) and one between group factor (label). Thus, all participants were exposed to each of the three functions, but each participant was exposed to just one of the three labels for each function. This reduced the design complexity associated with employing a fully repeated measures design (in which button location would need to be an additional variable). Participants were exposed to each label a total of six times (trials) enabling learnability to be explored. Within each trial participants were given five attempts to identify the correct button (enabling common mistakes to be identified). Thus there were a total of 18 trials for all functions (i.e., 3 functions x 6 trials) and 30 attempts over all trials for each function (i.e., 6 trials x 5 attempts). The presentation of the functions was fixed in the following order: 'Subtitles', 'Return to Sound and Vision' and 'Back to Previous Screen'. Fixed order was preferential to counterbalancing as the analyses aimed to explore labels within functions rather than cross-function comparisons. Further, fixed order kept fatigue and practice effects constant on performance across function.

The independent variable was button label (*A, B, C*), so each participant within a sample of 27 people would receive a different combination of the three labels (3 x 3 x 3). The dependent variables were accuracy (correct/incorrect) and reaction time (RT: in milliseconds). The RT data were cumulative – i.e., the RT to correct response (across attempts) was logged. If there was no correct response for a trial (across the five attempts), no RT was registered for that trial. Every button press (coded for button function) was recorded in the experimental trials so that common errors could be explored.

Virtual Remote Control and Procedure

A program was written to present the experiment (using Borland C++ Builder for Windows, Version 5.0). The graphic of the virtual remote control (see Figure 1) was constructed based on TEG-C recommendations for the required buttons on a DTT

remote control. The size of the buttons and the inter-button spaces conformed as closely as possible to guidelines produced by the Royal National Institute for the Blind (2003). Each function for which labels were evaluated in this Phase was placed on the same button throughout the experiment. In effect, only the labels of the three experimental buttons were varied within the experiment. Aside from a brief welcome and introductory note, the entire task was presented on an Iiyama INT3819Ts 15" LCD touchscreen.

Participants began with a series of 10 (unrecorded) practice trials which were unrelated in content to the experimental trials. The practice block aimed to (a) build participants' confidence with using a touchscreen in general, given that some would have never used one before, (b) familiarise them with the button size and halo catchment area for accurate button pressing, and (c) give them practice at positioning their finger over a red square presented at the bottom of the touch screen to initiate the onset of the experimental stimulus (i.e., the remote control). This was employed in an attempt to keep respondents' starting position for each trial as constant as possible so that measured reaction times could be meaningfully interpreted. In each of the practice trials the task was simply to press one of four buttons labelled A, B, C or D. Each time a button was pressed a tone was heard. As with the experimental trial, if the button press was correct, RT was recorded, the stimulus disappeared and the participant was moved on automatically to the next trial. If the button press was incorrect the stimulus remained on the screen until the correct button was chosen (up to five attempts).

Having completed the practice trial (which could be repeated if requested), participants were moved on to a familiarity task. The remote control replete with the specific configuration of button labels unique to that individual was presented on-screen for one minute. This aimed to reduce potentially large differences in exposure time to the remote control given likely individual difference in RTs on the experimental trials. For example, participant A might have taken longer to find the correct button on Trial 1 than Participant B and might have had the opportunity to more fully explore the remote control in this time. This could have potentially affected their RT for identifying the correct button for the next function. Before the remote control presentation, participants were simply instructed to familiarise themselves with the graphic rather than memorise it.

Participants were next moved on to the experimental trials. Each trial began with a function description. The function descriptions were written carefully to avoid using the critical labels in the descriptive text. The exception to this was 'Subtitles'. To describe this function without using the word 'subtitles' might have potentially confused participants – most people are now familiar with this function. Once they had read the instructions, participants pressed the red square to initiate the onset of presentation of the virtual remote control. On its presentation they were required to press the button they believed would activate the function described in the instruction they had just read. There were five attempts per trial and 18 trials in total. At the end of the experimental session all participants were thanked for their participation and fully debriefed. On average the full task took between 20 and 30 minutes to complete.

Participants

There were 27 participants aged 44 years or under. Fourteen were male and 13 were female with a mean age of 31.3 years (SD = 7.6, range = 18 – 44 years).

Another 27 participants, aged 45-84 years, have subsequently been run and data from these tests are currently being analysed. Data for this older sample will be available in a document prepared for the ITC to be released shortly, and a comparison of old and young responses will be published elsewhere.

Results

The resultant data were explored in a number of ways: (1) the number of people to correctly identify the label, (2) the number of attempts to make the correct response (or time out) on each trial, (3) types of errors, and (4) average time to make a correct response for each label (a) on each trial and (b) over all trials.

With 27 different combinations of the three labels for each function, nine people were exposed to the same label for any particular function.

Results for each dependent variable are reported for each function below. Most of the data is descriptive but analyses of variance are reported for the RT data over all trials.



Figure 1: Graphic of the Virtual Remote Control

Function 1: Subtitles

Number of people to correctly identify the label

For the 'Subtitles' function, almost everyone identified the correct label on each of the six trials (at some point across the five attempts). For both the 'subtitles' and 'subT' labels, all nine people correctly identified their respective buttons on each of the six trials. Only the subtitles symbol fared marginally less well – on the first trial, one person was unable to locate this button, but for every other trial, all participants were successful.

Number of attempts either to correct or time out

So how many attempts were needed to find the correct button? Across all trials, participants, on average, tended to find the correct button with one attempt – thus, all labels were relatively easy to identify quickly. However, there was some negligible variation between the labels. Participants needed fewer attempts to find the correct button on the first trial for the label 'subT' (1.22 attempts) than they did for 'subtitles' (1.33) or the symbol, which on average took nearly two attempts to identify (1.78). This pattern did not remain consistent across the six trials; by the second trial, the 'subtitles' label took just one attempt while for the other two labels it took slightly more

(1.11 for 'subT' and 1.22 for the symbol). On trial 3, 'subT' took the lead once more, but by trials 5 and 6 all three labels took just one attempt to find. On average over all trials, 'subtitles' and 'subT' needed the fewest number of attempts to correct – 1.07 each – while the symbol needed 1.2 attempts.

In terms of the absolute number of attempts required to first time correct across trials (out of a potential maximum of 30 attempts), most people (8 of 9) found the 'subT' label with one attempt, the other participant found it within three attempts. Perhaps surprisingly, just six of nine people found the label 'subtitles' on the first attempt; three people found it in two attempts. For the symbol, five people found it on the first attempt, three people on the second attempt, and one person took eight attempts (i.e., third attempt, second trial).

Types of errors

So what buttons were being mistaken for the 'Subtitles' function button? Six of the eight people that failed to correctly identify the 'Subtitles' button on the first attempt, chose 'Text' instead. Presumably these participants are basing their decision on the analogue TV model where subtitles are accessed first by entering text. Indeed one person used up four attempts by pressing the sequence, 'text'-'8'-'8'-'8'. Two people made the error of pressing the 'TV Guide – Electronic Programme Guide' function button (which took the form of an open book symbol on the remote control). Perhaps the association of 'book' with 'words'/printed text misled these participants to press this button.

Average time to make a correct response for each label on each trial

In terms of the average time taken to identify the correct label (across attempts) within each trial, 'subT' was most quickly identified on the first trial. It took about half as long to find this label than for the other two labels, which both took around six seconds to correctly locate. By the second trial, the differences between the times for each label were getting smaller with the symbol taking the longest to correctly identify. By trial 3, the labels 'subtitles' and 'subT' were virtually equivalent in time, with the symbol still lagging behind. Although this pattern changed somewhat on trial 4 ('subT' took the longest to identify), the pattern was consistent with earlier trials for trials 5 and 6. 'subT' was the only label for this function that was correctly identified in under a second by the final trial.

Average time to make a correct response for each label over all trials

This pattern of results was maintained when the data were explored in terms of the average time to identify the correct label across all trials. Participants were able to learn each label with time (each trial), however, 'subT' was relatively easier to learn on average, than either 'subtitles' or the symbol. The label 'subT' took an average 1.5 seconds to find over all trials (n = 52) compared with 'subtitles' (1.9 seconds, n = 52) and the symbol (2.1 seconds; n = 51). A one-way ANOVA revealed no statistically significant differences between the labels in these mean times ($F_{(2,154)} = 1.88$; ns) (see Figure 2). This may be due to low power because of the small sample size available here. Throughout the results section of this paper, this is a consideration of which the reader should be aware.

(The results presented in Figure 2 exclude 6 outliers i.e., $> 2.5 \times S.D. + \text{mean}$, who responded in excess of 10 seconds. Before deletion of these cases, the mean times to correct across trials were 2.7 for the subtitles symbol (n = 53), 2.5 for 'subtitles' (n = 54) and 1.9 for 'subT' (n = 54).)

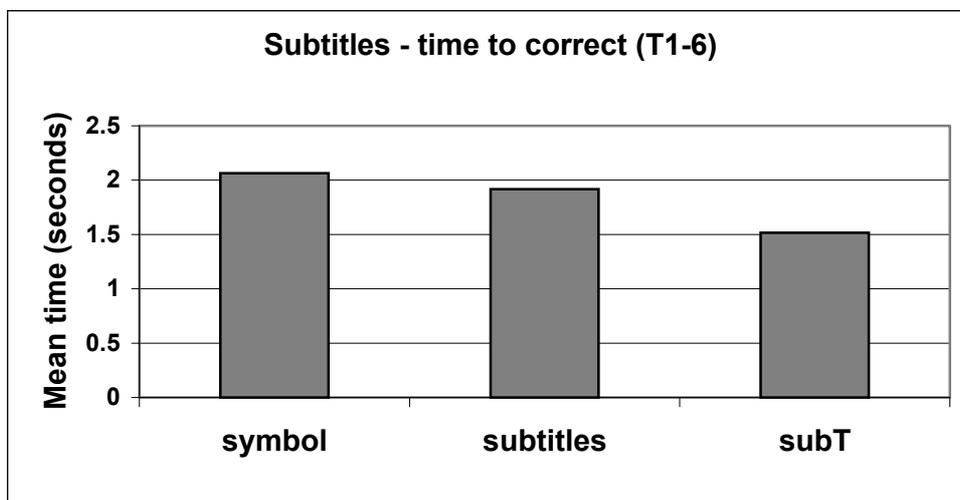


Figure 2: Mean time to identify the correct Subtitles label over all trials

Function 2: Return to Sound and Vision (from any Interactive Application)

Number of people to correctly identify the label

Although all participants correctly identified their respective labels by trials 5 and 6, there was some variation in the earlier trials. As predicted on the basis of the results

from the subjective preferences in Phase 3, 'out' fared least well – seven people (of nine) made a correct button press on the first trial, rising to eight by the second trial, but dropping again to seven people by the third trial. Nevertheless, by the fourth trial eight people were successful again and there was 100% success for the fifth and sixth trials. 'TV' and 'return' also showed similar variation, indicating some difficulty in remembering the correct label. 'TV' started with just one person unable to identify the correct button, but by trial 2, all participants made the correct choice. This dropped to eight people again for the third trial. By the fourth trial and for the remaining two trials, everyone correctly identified 'TV'. Finally, 'return' started well – for the first two trials, all participants found the correct button. But on trial 3, one person failed to correctly identify 'return'. Nevertheless, on the last three trials, 100% of participants found the correct button.

Number of attempts either to correct or time out

Overall people tended to have more difficulty with this function than with 'Subtitles'. On the first trial all participants, irrespective of label, took an average of around 2 attempts to find this button or time out. As anticipated, the label 'out' on average took more attempts (2.56), than 'TV' (2) or 'return' (1.78). Across the trials, 'out' also showed considerable variation and always took an average of around two attempts to find with each trial (except for trial 6 when it took on average 1.22 attempts). This indicates that 'out' was less memorable and more difficult to learn. Interestingly, 'return' also showed some variation – while it outperformed the other labels on the first trial and although the average number of attempts generally declined over trials, it increased on trials 3 and 6. It appears that 'TV' showed the most promising pattern of results; by trial 5 all participants were getting it right first time.

In terms of how many absolute attempts it took to get it right across trials, five (out of nine) people made the correct choice on the first attempt for both 'TV' and 'return', while only three people correctly identified 'out' on the first try. For 'TV' one person needed two attempts, two people needed three attempts, and one person needed six attempts to find this button. For 'return', three people found it in two tries, and one in six tries. And for 'out', two people needed two attempts, two people needed three attempts, one person needed six attempts, and another person needed 25 attempts to find the right button (i.e., they did not correctly identify the button until the last attempt of the fifth trial).

Types of errors

So what common mistakes were made? Six out of the fourteen people who failed to locate the correct button on the first attempt made the mistake of pressing the button corresponding to the 'Back to Previous Screen' function. Two people first chose the 'Text' function, two selected the 'TV Guide – Electronic Programme Guide' function button (an open book symbol), and one person incorrectly selected each of the following: left pointing navigation button, 'menu', 'info' and the 'Subtitles' function.

Average time to make a correct response for each label on each trial

In terms of the average time taken to identify the correct label (across attempts) within each trial (learnability), on trial 1 'return' was the most quickly identified label, taking nearly five seconds less time to locate than for either 'out' or 'TV'. By trial 2, 'return' in fact took slightly longer than previously (8.2 seconds) while 'TV' and 'out' were much more speedily identified in around two seconds each. This was largely due to an outlier for the 'return' label – a participant who was unable to locate 'return' on the first trial, but who took 28.5 seconds on the second trial – about 13 seconds longer than the next slowest participant. When this outlier was removed, the mean time for 'return' on the second trial was 5.67 seconds. The pattern shifted over the trials. By trial 6, 'TV' was the most quickly identified label, taking about 1 second to find.

Average time to make a correct response for each label over all trials

In terms of general learnability (time to correct across trials), 'TV' emerged as the winner. 'TV' took an average of 2.4 seconds to locate ($n = 50$), compared with 3.6 for 'out' ($n = 47$) and 4.2 for 'return' ($n = 51$) (see Figure 3). One-way ANOVA suggested that these differences approached significance ($F_{(2,147)} = 2.7$; $p = 0.07$). Specifically, post-hoc tests (Games Howell correction – equal variances not assumed) revealed that the difference between 'TV' and 'return' approached significance ($p = 0.052$).

(The results presented in Figure 3 exclude four outliers whose scores fell above 2.5 SDs from the mean – i.e. greater than 19.6 seconds. Before deletion of these cases, the mean times to correct (across trials) were 3.6 seconds for 'TV' ($n = 52$; corrected = 2.4), 4.1 for 'out' ($n = 48$; corrected = 3.6) and 4.6 for 'return' ($n = 52$; corrected = 4.2).)

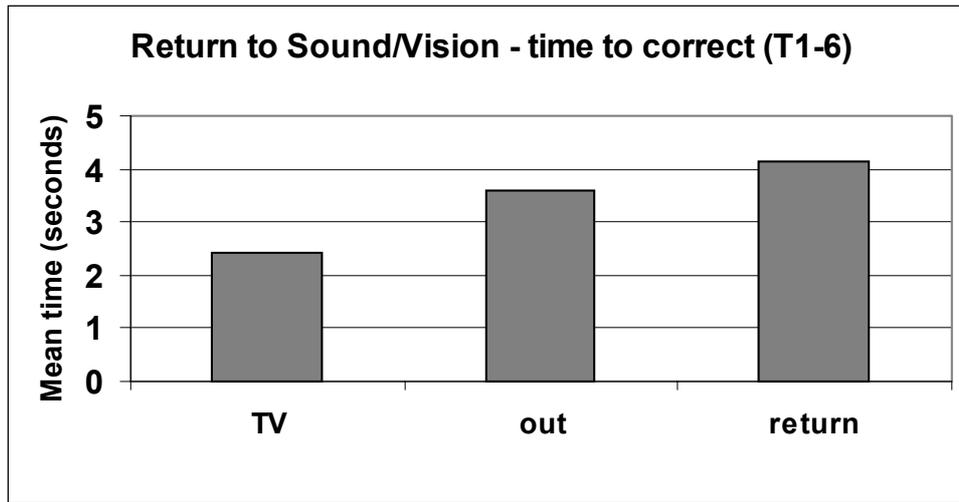


Figure 3: Mean time to identify the correct Return to Sound and Vision label over all trials (learnability).

Function 3: Back to Previous Screen

Number of people to correctly identify the label

The results were most revealing for the 'Back to Previous Screen' function. The label 'back' was correctly identified by all nine participants on every trial. In comparison, 'back up' needed two trials before everyone made the correct choice (and continued to do so for the remaining trials). More people had problems with the label 'cancel'. On the first trial, just four people identified this button, rising to six people for the next two trials, then seven people for trials 4 and 5. It was only on the sixth trial that all nine people correctly located this button.

Number of attempts either to correct or time out

Of all the functions the biggest variation between labels in the number of attempts to correct was found for 'Back to Previous Screen'. The least popular label subjectively, 'cancel', on average took participants many more attempts to find than either of the other two labels. On the first trial, 'cancel' took an average of four attempts to correctly locate and press. Furthermore, although there was a steady decrease in the number of attempts with each trial for 'cancel' (indicating learnability), even by trial 6 the average number of attempts was around two (1.78). The label 'back up' required an average of around three attempts to find on the first trial (2.78). Again, this label showed a steady decrease in the number of attempts needed with each trial. But in contrast with 'cancel', for 'back up' all participants identified this label correctly on the first attempt by trial 4. Clearly, this label was easy to learn. However, the label 'back' was the clear winner here. It consistently needed around just one attempt to correctly identify and press from the first trial right through to the (final) sixth trial. Over all trials,

on average 'back' needed 1.09 attempts, compared with 1.46 attempts for 'back up' and 2.56 attempts for 'cancel'.

In terms of the number of attempts to first time correct, nobody was able to find the 'cancel' button on the first attempt. Three people were able to find it in two attempts, and one on the fifth attempt. Five people took seven attempts or more to find the label 'cancel'. In fact, one person found the correct button only on the last attempt of the sixth trial (i.e., 30 attempts to correct). For 'back-up', there was also some difficulty. While four people managed to correctly locate this button on the first attempt, one person did it in three attempts, two people in four attempts, one in nine attempts, and another in 13 attempts. The label 'back' was most easy to identify in the fewest number of attempts. Seven out of the nine participants found 'back' on the first attempt. One person needed two attempts, and another needed three attempts.

Types of errors

In terms of common mistakes, 10 of the 16 people who failed to choose correctly on the first attempt instead chose the left pointing navigation button. Two people pressed the 'TV Guide – Electronic Programme Guide' function button (an open book symbol) and one person each mis-selected one of the following: right pointing navigation button, 'Text', the 'Return to Sound and Vision' function button, and the downward pointing navigation button. That the open book symbol ('TV Guide – Electronic Programme Guide' function) was mistakenly pressed for this as well as the other two functions suggests that the meaning of this symbol is not entirely clear to participants. However, their readiness to try this button might indicate that it is perceived as a 'friendly', 'no harm done' function. It is interesting that the navigation buttons were a common error. Clearly, they were an intuitive option for many participants, and actually require no label whatsoever.

Average time to make a correct response for each label on each trial

First, in terms of the average time taken to identify the correct label (across attempts) within each trial, the label 'back' was the most quickly located label on the first trial, taking about 5 seconds to find – three times as quick than for 'back up' (15.2 seconds) and almost five times faster than for 'cancel' (23.8 seconds). This pattern was similar for the next two trials, but with the differences in times between the labels getting smaller. By the fourth trial, 'back up' was the most quickly identified label, and this continued for the remaining trials. The differences between 'back' and 'back up' were

negligible, however. 'Cancel' was also more speedily identified with each trial. On trial 6, the mean time for 'cancel' was much larger than for the three previous trials for this label (19.3 seconds). This, however, was due to an outlier who took 135 seconds to find this label on the final trial. When this outlier was removed, the mean for trial 6 for 'cancel' dropped to 4.76 seconds.

Average time to make a correct response for each label over all trials

All labels were faster to identify over progressive trials, indicating learnability. On average over all trials 'back' took 2.1 seconds (n = 54). This was faster than for 'back up' (3.4 seconds; n = 50) and almost three times faster than for 'cancel' (5.9 seconds; n = 36) (see Figure 4). A one-way ANOVA revealed that there was a significant difference in time to correct (across all trials) between the three labels ($F_{(2,139)} = 3.46$; $p < 0.05$). Post-hoc tests (Games-Howell correction – equal variances not assumed) revealed that the difference between 'back' and 'cancel' approached significance ($p = 0.053$) but there were no other significant differences.

(The results presented in Figure 4 exclude four outliers (time scores in excess of 2.5 S.D.s away from the mean i.e., in excess of 40.7 seconds). Prior to deletion of these outliers, the values for time to correct (across all trials) were 4.2 seconds for 'back up' (n = 51) and 11.3 seconds for 'cancel' (n = 39). There were no outliers for the 'back' function.).

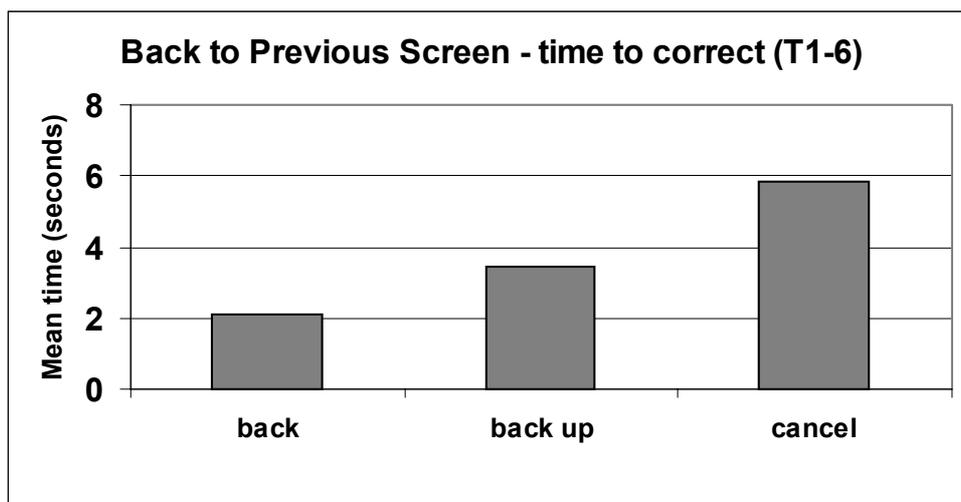


Figure 4: Mean time (in seconds) to identify the correct over all trials (learnability)

3. Discussion and Conclusion

Generally there was good correspondence between the subjective preference data and behavioural performance data, although the relative differences between the labels across the subjective and behavioural results were not always of equivalent magnitude. These results are reassuring and demonstrate that subjective preferences can be a good indicator of behavioural performance. Further, subjective preference measures are quicker, cheaper and easier to collect than are behavioural measures. However, the behavioural data are important to consider. For example, how many times will a viewer be willing to press the wrong button and get an unexpected system response before giving up on trying to use a service, and telling all his/her friends that it's impossible to use?

In terms of the behavioural data, the results demonstrate that for the 'Subtitles' function, whilst the label 'subtitles' was the (subjective) favourite label 'subT' outperformed 'subtitles' behaviourally (although statistically there was no significant difference between the times to locate these labels). This finding might relate to the label size – perhaps it was more difficult for subjects to read this text when it almost fills the button. Further, the 'subT' label might be more eye-catching because of the unusual capital letter at the end. It is also possible that younger people are more familiar and confident with using ad-hoc abbreviations from mobile phone text messaging, the use of which is particularly popular among younger age groups. It will be interesting to compare this result with data from the older age sample. The internationally recognised subtitles symbol was responded to least well, as predicted on the basis of the results from phase 3. It was interesting to note, however, that a common error in locating the 'Subtitles' function was access via the analogue TV route – in the UK, analogue TV viewers access subtitles by pressing text, and then 888. This suggests that in addition to having a dedicated subtitles button, designers should allow users to access subtitles via an option in the 'text' route - as this is what they are used to. Clearly, as a general rule, maximising the consistency between operating a novel technology (digital TV) and what people are used to (with analogue TV) is good practice, where possible, and will result in digital TV being easier to use for everyone.

For the 'Return to Sound and Vision' function, phase 3 showed that 'return' and 'TV' were close contenders for favourite, while 'out' was the least popular. The behavioural results for these labels demonstrated that 'out' caused participants the most difficulty.

Compared to the other labels, not only did fewer people actually identify 'out' on each trial, but they also needed more attempts to locate it correctly. The behavioural results were particularly revealing in identifying the relative usefulness between 'return' and 'TV'. Over the trials there was just one occasion where 'return' was not identified (at any point over the 5 attempts per trial) compared with two occasions for 'TV'. Further, 'return' needed marginally fewer attempts on average to find on each trial. However, on average (across all trials) participants responded to 'TV' more quickly than they did to 'return' and showed good learning for 'TV' over the trials.

While 'return' outperformed 'TV' in some analyses, during debriefing a few participants commented on the slightly ambiguous meaning of the term 'return'. Further, it was observed that the correct button for the functions 'Back to Previous Screen' and 'Return to Sound and Vision' were often confused by participants (e.g., pressing 'back' instead of 'return'), possibly because of the repetitive nature of the task. This confusion was exacerbated when the two terms for the different functions were more similar semantically and was reduced when the label 'TV' was used. It appears that, all results considered, 'TV' was the winning label.

Finally, for 'Back to Previous Screen', the behavioural results lent support to the subjective data. The label 'cancel' consistently performed least well of the three labels, as predicted on the basis of the results from phase 3. While 'back up' performed slightly less well to start, once learned, it was easily remembered. The label 'back' consistently performed well, as predicted on the basis of the results from phase 3. Indeed, the time to correct results showed that 'back' was responded to significantly more quickly than was 'cancel'.

A few observations are worthy of note in interpreting the results from this work. First, participants were often observed to be circling their finger around the centre of the remote control where the 'OK' and navigation buttons were located. The most interesting shapes and button spacings were located in this region suggesting that button location is an area worthy of future research.

Second, the observation that the 'Back to Previous Screen' and 'Return to Sound and Vision' functions were sometimes confused with each other, highlights the importance of context in terms of the labels given to the other buttons on a remote control. That is,

where a function may be ambiguous, function meaning is derived by comparison with possible alternatives.

Further, while it was aimed to present the behavioural study in an ecologically valid way, in everyday situations there are usually multiple means of accessing the same function - a reality that was not accounted for in this study. For instance, in many circumstances, a series of 'back' presses should eventually remove the user from an interactive application (as well as in one button press using the 'Return to Sound and Vision' function). In addition, in everyday remote control use, buttons are usually identified based on the users' understanding of what function they would like to activate, either directly when faced with the remote control, or indirectly when they are faced with options within menus and submenus. This is in contrast to the behavioural paradigm employed here where a description of a function was provided for them.

Thus, context (the influence of surrounding buttons and what they have been labelled) is clearly an important influence on people's decision making in this task. While this was kept standard (other than for the critical buttons which varied from person to person), and conformed to TEG-C recommendations, the extent to which the non-target buttons and labels used here affected performance on the critical labels is unclear.

The sequential methodology employed here, the design of the behavioural paradigm and the time scale available for this project made it necessary to reduce the number of functions and labels being tested at each phase. However, for a complete understanding of the intuitiveness and learnability of labels (for both existing and novel functions) on a given remote control, it would be necessary to explore all suggested function labels and possible label configurations.

The study used both qualitative and quantitative research methodologies in developing and answering the research question, enabling exploratory inquiry to develop into structured methods. Both users and 'experts' (from industry) participated in the study enabling usability (subjective opinion and objective behaviour) and pragmatic concerns of manufacturers to be addressed. The correspondence between phase 3 (subjective preference) and phase 4 (behavioural performance) was explored indirectly by comparing the patterns of results using both methods i.e., using

comparative estimations of 'best' through 'worst' label. However, there was no concrete benchmark by which to judge 'good' compared with 'bad' performance. Further, the labels were compared within-function – some functions are generally easier to understand and use than others. The results suggest that in general, subjective opinion may be a good, albeit rough, indicator of behaviour, though the differences between performance on the labels were of a lesser magnitude than the differences in subjective opinion. There was no evidence of very poor correspondence between phases 3 and 4 although it is possible that there might have been had all functions and labels been tested. Examples of poor subjective-behavioural correspondence would be of interest to designers. Subjective preference is an important criterion in usability research – it can be important to give people what they think they want as users' overt evaluations can sometimes override the performance benefits. *Perceptions* of ease of use are critical. However, they are not a direct substitute for behavioural measures, particularly more ecologically valid paradigms such as that employed here. Thus, whilst this study has shown that subjective preferences can be good indicators of behavioural performance, for new products and services in development, we recommend using a combination of both subjective and objective measures.

A final key point is when evaluations of the type reported here should be conducted. In terms of high support costs, high return rates, and bad word of mouth for brands (and product categories) it is potentially very expensive to design products without sufficient regard to their ease of use for their target users. Measuring target users' subjective and behavioural responses to software based mock ups of new products and services in development (using ecologically valid paradigms) is most definitely a worthwhile investment in the future success of these products.

4. References

- Freeman, J., & Lessiter, J. (December 2001). Easy TV 2001 Report: ITC-UsE Ease of Use and Knowledge of Digital and Interactive Television – Results. [online] <http://www.itc.org.uk>
- Freeman, J., Lessiter, J., Williams, A., & Harrison, D. (January 2003). Easy TV 2002 Report. [online] <http://www.itc.org.uk>

Gill, J., & Perera, S. (2003). Accessible Universal Design of Interactive Digital Television. [online] <http://www.tiresias.org/reports/brighton.htm>

Royal National Institute for the Blind (2003). Guidelines for the Design of Accessible Information and Communication Technology Systems: Technology – Keypads. [online] www.tiresias.org/guidelines/keys.htm.