Measuring acceptable input: What is “good enough”?

Simeon Keates

University of Greenwich,
Medway Campus, Central Avenue,
Chatham Maritime,
Kent ME4 4TB
United Kingdom
E-mail: s.keates@gre.ac.uk

Abstract

Many new assistive input systems developed to meet the needs of users with functional impairments fail to make it out of the research laboratory and into regular use by the intended end users. This paper examines some of the reasons for this failure and focuses particularly on whether the developers of such systems are using the correct metrics and approaches for evaluating the functional and social attributes of the input systems they are designing. This paper further focuses on the importance of benchmarking new assistive input systems against baseline measures of useful interaction rates that take allowance of factors such as input success/recognition rate, error rate, correction effort and input time. By addressing each of these measures, a more complete understanding of whether an input system is practically and functionally acceptable can be obtained and design guidance for developers is provided.

Keywords

Interaction rate, universal access, HCI, input technologies, error rate, assistive technologies, acceptability

1 Introduction

Much of the research into Universal Access, both past and present, has focused on the development of new and innovative assistive input device and interface design technologies for users with functional impairments. It is widely accepted that the traditional keyboard and mouse input arrangement does not serve those with a range of functional impairments well [1]. New technologies are being introduced that do not rely so heavily on the traditional mouse and keyboard set-up [e.g. 2], but are still typically not being developed with users with functional impairments in mind.

A person with severe vision impairment will experience significant difficulties in using a mouse, not least because the feedback on the position of the cursor on the screen is invariably visual only. Similarly, users with motor impairments will typically experience comparable levels of difficulty, because of the challenges presented in generating the quality of limb and digit control usually required to position a mouse, click on its buttons or type on a keyboard [3]. Consequently, many researchers have taken the view that perhaps a new input device / user interface arrangement [e.g., 4] or a re-design of the device/interface [e.g., 5] may alleviate or remedy the difficulties faced by many such users. Tablets, for example, do not use keyboards or mice/pointers in the same way as, say, a laptop or desktop, but on-screen keyboards and direct touch interfaces still present major accessibility challenges to users with vision and motor impairments [6], as well as older adults [7].

However, while the motivation for developing new assistive input and interaction technologies is clear, the success of such devices has been mixed. It is still a common problem that many of the new technologies developed rarely progress beyond the research laboratory. Of those that do, many end up simply collecting dust on shelves, never really used to the extent anticipated by their developers [8].

There are many reasons why individual assistive input technologies suffer this fate, although there are a few that are reliably useful indicators of the likely success or otherwise of such developments. Jakob Nielsen, for example, has identified that the success of a product depends on it meeting both practical/functional acceptability and social acceptability criteria [9]. He defines practical acceptability as including factors

such as cost, reliability, utility/functionality and usability. Social acceptability considers factors such as brand identity, stigma, etc. and research has been undertaken to explore how these factors can be investigated in a universal access context [10].

There is a large body of work looking at usability theory and overall acceptability of products and systems. Of particular interest in the context of this paper is the challenge of establishing whether the practical acceptability offered by input systems has genuinely been met. While it is straightforward to obtain some measures of functional acceptability through even quite short user trials, developers typically look at only a subset of the interaction when evaluating their new systems. The challenge is to identify a more complete set of metrics that are practical to evaluate.

It is accepted that one of the principal reasons for the failure of the uptake of these new solutions is that their development has typically focused on the functional/technical issues, i.e., getting the solution to work, often to the detriment of the softer/social issues, i.e., does it meet the wants, needs and/or aspirations of the users [11]. Indeed, this can often be considered the "irony" of universal access research where researchers looking to develop improved interfaces can sometimes find themselves side-tracked into developing new hardware technologies first. This "irony" is not the fault of the researchers, it is an unintended consequence of how funding bodies typically structure their calls for proposals and measure the success, or otherwise, of their outcomes. Many funding bodies fund up to the point where a prototype has been developed, though then assume that a commercial partner will take over and push the product out into the marketplace. The funding bodies typically stop monitoring progress at that point. However, a research prototype is usually far from market-ready and significant further investment is often required to increase the technology-readiness level. It is rare that funding streams are available to support that next phase of development. This problem is not new, as similar opinions have been aired almost 20 years ago in relation to the development of rehabilitation robotics [12] funded through EU TIDE projects.

For example, if there is a funding call for, say, ambient intelligent environments, that funding can be leveraged more successfully to investigate particular features of interaction for users with severe motor impairments under the guise of ensuring that

the ambient intelligent environments are "accessible for all" than perhaps a direct funding proposal to look at the interface issues alone may be. The downside, though, is that the team then needs to dedicate time to the hardware development, which although not a problem directly, history has shown can tend to expand and end up dominating the research effort. The consequence of focusing on the technological development, if not managed appropriately, is that the user interface and user-centred design activities are often relegated to later in the development process, contrary to all the published literature and guidance. Hence, research that was intended to look at improving the interface often ends up failing to achieve the promised advances for the same age-old reasons – and the unfortunate "irony" of much universal access research and funding being focused on getting to the point where the core research becomes possible. Consequently, much of the data presented in this paper is derived from very traditional input configurations, such as the keyboard and mouse or gestures, where the technology is mature and stable enough to allow sufficiently detailed analysis of user interactions since research effort does not need to be focused as much on the development of the input technologies.

It is necessary to recognize that a failure to meet the practical acceptability criteria will also translate to a failure of the product or system to succeed in the real world. For assistive input systems, assuming that the project is correctly managed to avoid excessive focus on technological development issues, one of the major difficulties has been that the functional aspect of the development often only considers a narrow part of the interaction process as the metric of success. In many cases, this is usually input recognition rate [e.g. 13]. This paper explores a more complete approach to evaluating interaction and assessing whether an input system delivers a genuinely acceptable solution for users. Although the data presented is necessarily based on specific and somewhat mature input technologies, the general principles are transferrable to newer and emerging technologies.

2 Functional Impairments and Universal Access

To begin considering methods of evaluating the effectiveness and acceptability of input systems for users with functional impairments, it is necessary to begin with a

brief summary of the major categories of functional impairments and how they affect human-computer interaction.

There are several approaches to categorizing types of functional impairment that can be used by designers and developers of new input systems. One of the most straightforward was inspired by the work of Card, Moran and Newell on the Model Human Processor [14]. Effectively, they proposed a model of interaction that consists of three elements:

$$\text{Total time} = x\,\tau_p + y\,\tau_c + z\,\tau_m \qquad (1)$$

In this equation, x, y and z are integers and τ_p, τ_c and τ_m correspond to the times for single occurrences of the perceptual, cognitive and motor functions respectively. It is possible to categorize impairments along these lines of functionality.

Perceptual impairments are those that affect a user's ability to perceive the state of the world around them and are principally focused on the five senses. In the case of computer access, the human senses of most interest are vision and hearing [15]. Indeed, vision impairments have historically received arguably the lion's share of research effort and also have the most successful assistive technologies to facilitate better interaction, with products such as JAWS achieving strong market positions [16]. Blindness and low vision present challenges with most stages of human-computer interaction, from input actions, such as text entry and cursor control, to perceiving output, such as reading text on a screen or interpreting a figure or diagram.

Cognitive impairments are those that affect the user's ability to understand or respond to the state of the world around them. Such impairments can include memory loss or reduction, learning and communication difficulties and executive function limitations [17]. It is often argued that cognitive impairments are the most "hidden" ones, since their presence is often more difficult to identify and, once identified, to also diagnose. However, they are beginning to be researched more frequently [e.g. 18] than, say, 10 years ago. Typical solutions can include personalized diaries and reminders for medication and other reminders, assistive word processors for help with typing and dialogue structures, etc. More innovative solutions include emotion and affective state

recognition to assist people with Asperger's and forms of autism [19] and also deep question and answer systems, such as IBM Watson [20].

Finally, motor impairments can create difficulties with both text entry and cursor control in a typical computer interaction scenario [1]. Symptoms such as tremor, spasm, restricted range of motion and weakened muscles can make both gross and fine motor control a challenge [1].

Text entry assistance typically focuses on making keyboards more accessible through physical assistance, for example adding key guards, or using "soft" on-screen keyboards or replacements, such as Dasher [21]. On-screen, soft keyboards are usually activated by a dwell time function (in the case of a cursor control replacement system) or some form of binary switch / scanning combination [8].

Cursor assistance can be in the form of adapted mouse replacement devices, such as tablets or specially designed mice/joysticks/trackballs [1]. One area of particular promise is that of haptic assistance, such as through the addition of "gravity" to on-screen targets [22]. Other approaches include adapting or altering the processing of the cursor input stream to make targets more "sticky" by slowing the cursor down over the targets or by fixing mouse button activation to the location of the button down event, not the button up one [23]. More radical solutions involve changing the input paradigm from the usual windows/icons approach to that of using gestures for the input [24], for example.

As can be seen, there are many forms of functional impairments that can affect human-computer interaction adversely and present specific challenges to particular users. There are also many forms of potential assistance, each of which offer their own particular combination of strengths and weaknesses. As discussed earlier, not all of these assistive solutions are successful in the wild, so the question then becomes whether there are more effective methods for identifying or predicting whether a particular solution has a genuine chance of successful adoption by users in real world circumstances.

As regards determining the social acceptability of a new technology or product, approaches such as focus groups, user evaluations, etc., would usually be used [25]. These methods are generally well understood and widely accepted. However, there is less of a consensus on methods of evaluating the practical acceptability of novel interaction technologies.

3 Defining "Acceptable" Interaction for Universal Access

Most research papers addressing the development of novel input systems or interaction paradigms usually focus on only one or two measures of success, principally the rate of successful completion of a specified task, such as clicking on a target or producing a particular gesture that is recognized correctly by the computer. While clearly a very important measure, focusing on this metric only can lead to an exaggerated view of the efficacy of the new input system/interface. There are other important factors to consider, such as the definition of usability used by ISO [26]:

- Efficiency, i.e. the time taken and effort expended to complete a task;
- Effectiveness, i.e. the ability to complete the task;
- Satisfaction, i.e. user contentedness with the interaction.

Using these definitions, satisfaction is typically measured through user surveys, interviews, questionnaires, etc., after completing a series of tasks using the new technology [27].

Efficiency is usually calculated by looking at the task completion rates and times. In most cases of developing new input systems for users with functional impairments, recognition rate is the measure used most commonly. Task completion rates and the time to complete tasks are sometimes reported, though not often, and certainly not in all research papers.

Measuring effectiveness involves looking at error rates and effort expended to correct for any errors that occur as well as proportion of tasks completed [27]. However, while research papers addressing the development of assistive input systems that include some form of user evaluation with the prototype system usually include a

summary of task completion times (i.e. a variant of the efficiency metric above) and task completion rates (i.e. a partial treatment of the effectiveness metric), it is less common to find an exploration of the frequency of errors. It is even less common to find an analysis of the impact of those errors, with some experimental designs not even recognizing the presence of errors.

Even in the comparatively rare instances where such analyses exist, it can be argued that the final piece of the jigsaw is still missing – i.e. a comparison with an accepted baseline measure. Fundamentally, even where the developers do such analysis, they often fail to reflect on whether the assistive input system that they have developed meets an acceptable level of interaction. It is all well and good to say that it takes x seconds to complete a task, with an error rate of y%, however the real question is whether those task completion and error rates are acceptable to the intended end users [28].

User satisfaction is also rarely considered explicitly in the development of new input systems for users with functional impairments. Some authors do use standardised measures of task load, such as the NASA TLX questionnaire [e.g. 29], though it is rare to see a discussion with the users about whether they prefer the new system to any other system they may have used. Where such questions are asked, the authors rarely control for the different levels of exposure between the systems, i.e. they do not typically seem to compensate for the fact that one system may be very familiar to the users, whereas the other, by definition since the research is about a novel input system, would be very new to them.

3.1 An Approach to Evaluating "Acceptable" Interaction

If a new input system is to be considered acceptable to the end user and also likely to be used "in the wild," it needs to be a number of straightforward targets. For example, one obvious question to ask is:

- *Does this new assistive input system equal or outperform the other systems available to the end users?*

If the answer to this question is negative then that immediately casts doubt upon the likely successful adoption of the system being developed by users outside of the research laboratory. Fundamentally, if users can obtain better interaction rates using an existing, and most likely proven, assistive input system, then they are less likely to wish to switch to a new or different one. Even if the recognition rates appear to be good, for example 95% or higher, if the existing input system used by the user offers, say, 98% recognition, then there is little reason for the user to consider changing to the new system based on that metric.

For users with more severe impairments there may not be a suitable or practical input system readily available. However, in all but the most extreme cases, some form of input is usually possible through the use of simple binary, i.e. on/off, switches and a scanning on-screen keyboard. Consequently, it can be argued that the very minimum target for user acceptance of a new assistive input system is that it should at least outperform the scanning/binary switch input approach. Ideally, given the effort typically taken to learn and master a new input system, it should outperform any existing available system by some distance.

Even where the answer to the question above is positive, there are further questions to be asked, for example:

- *Does this new assistive input system meet the full needs, wants and aspirations of the end users?*

Where the first question focuses on the practical acceptability of the new input system, the second focuses on the social acceptability. Once the answers to these questions have been derived, it is possible to ask a third one, specifically:

- *Is this new assistive input system good enough?*

Answering a question of this type is not straightforward, as a quick read of any good book on usability makes clear. In the case of human-computer interaction there are a few principal input metrics that need to be considered: text input, cursor input and

overall interaction rate. A further metric also needs to be explored: cognitive load on the user.

3.2 Measuring Text Input

Text input has been studied in great depth [e.g. 30, 31] and is typically reported in terms of words per minute [e.g. 7]. It may also be reported as characters per minute, if that is a more meaningful metric, such as when typing rates are unusually slow or where a more detailed analysis is required [32].
However, defining a "word" is not straightforward. Many approaches simply assume that a word is 5 characters in length, with a following space implicitly (5 characters) or explicitly (5+1 characters) associated with it. In many modern systems, the impact of word prediction systems needs to be considered. It is not clear how often users need to actually enter all 5 characters to make a word when a predictive system is also being used, thus raising a question over the calculations made using the 5 or 5+1 assumptions.

There is a choice to be made over how to handle errors. Some researchers simply choose to ignore that errors may exist, e.g. by not supporting or allowing error correction in the design of the experiment. Others remove words with errors in them from the data analysis. Neither of these can be considered ideal solutions when looking at users with motor impairments where errors will most often carry a significant correction penalty, i.e. the amount of effort required to correct any errors will be non-trivial, and also where the frequency of errors can be expected to be significant.

Where errors are identified, they are typically reported through metrics that capture deviations from the expected minimum, error-free input, such as Mean String Distance (MSD) or Keystrokes per Character (KSPC) [33]:

$$MSD \approx \frac{INF}{C + INF} * 100\% \qquad (2)$$

$$MSD \approx \frac{C + INF + IF + F}{C + INF} \qquad (3)$$

where: INF = Incorrect and Not Fixed character entries,

IF = Incorrect but Fixed,

F = Fixing non-character entries (e.g. a backspace or other edit function); and,

C = Correct character entries.

Other measures are possible [33], though are not used as often as MSD and KSPC.

3.3 Measuring Cursor Input

The most common approach to measuring cursor input is to use a Fitts' Law type experiment. Fitts' Law has undergone a number of modifications since first proposed, and the Shannon formulation is one of most commonly used [e.g. 34]:

$$Movement\_Time = a + b * ID \quad (4)$$

where *a* and *b* are constants and the Index of Difficulty (*ID*) is:

$$\mathrm{ID} = log_2\left(\frac{D}{\mathrm{W}} + 1\right) \quad (5)$$

in which *D* is the distance travelled towards the target and *W* is the width of the target along the direction of travel.

Although experiments have confirmed that Fitts' Law can be applied to users with motor impairments, there is again little explicit handling of errors. A more sophisticated set of cursor measures has been developed to look at the detail of the quality of cursor control [35] and these measures have been applied successfully to examine the quality of cursor control for users with severe motor impairments [36]. Again, though, while these measures can tell a lot about what is happening to the cursor input, they do not necessarily help researchers and designers determine if the quality of the input is sufficiently good. There is a clear need for a baseline measure to compare against.

3.4 Measuring "Useful" Interaction Rate

As can be seen from the discussion above, there are many ways of examining the details of human-computer interaction. However, while those methods may make good research tools, they do not typically answer the question raised earlier, specifically: is the input system good enough?

To answer this question succinctly, a simple metric needs to be considered, one that can help a developer or researcher know immediately if the new system is operating in the correct ballpark. A likely candidate for such a measure is the bit rate of useful information transfer between the user and the computer utilizing the assistive input system. Fitts proposed such a measure, throughput, calculated as:

$$Throughput = \frac{ID_e}{MT} \quad (6)$$

where MT is the movement time described above and ID_e is the effective index of difficulty:

$$ID_e = log_2\left(\frac{D}{W_e} + 1\right) \quad (7)$$

based on the effective width, W_e and the initial distance to the target, D. However, while this serves as a very useful and popular metric in most circumstances [e.g. 37], it arguably does not take into account the full impact of the presence of errors in typical interaction patterns for users with more severe functional impairments. A modified version of this measure is required.

An example of how such a measure can be generated is illustrated by a gesture recognition system [38]. In that experiment, users were able to generate a range of possible gestures (the vocabulary). Rather than using a simple recognition rate, a scoring system was implemented where correctly recognized gestures were scored as +1, non-recognized gestures were scored as a 0 or null return and misrecognized gestures were scored as -1 to reflect that a corrective action would be needed to fix the error. The overall input samples gathered from each user were then normalized and scaled to a range of -100 to +100 to remove any data collection issues, such as incomplete task or data sets.

That score was then combined with the vocabulary size and the time taken to produce and recognize each gesture into a single measure, the bit rate of useful information transfer between the user and the system:

$$Bit\_rate = \frac{\log_2(Vocabulary\_size) * \frac{Score}{100}}{Time\_taken} \quad (8)$$

It can be seen from the formulation of equation (8) that a system scoring 0 or less will not generate any useful bit rate since the user will be permanently trying to correct incorrect inputs, which is intuitively correct. Figure 1 shows the scores obtained for

single mode gesture recognition from [38], for 3 or 6 gestures made by the user's head.

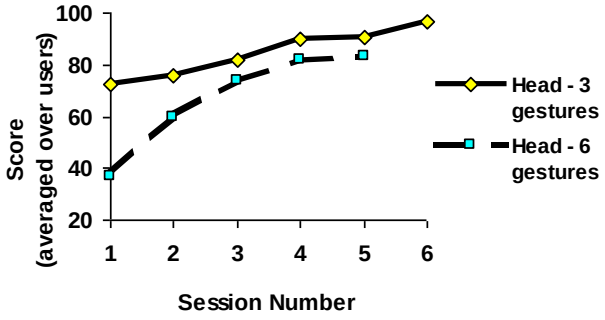

Figure 1. The interaction scores obtained for single mode gesture recognition from [30]

It is worth noting that the scores could be modified from the +1, 0, -1 values used. While retaining +1 makes sense for a successful input, it could be argued that a non-recognised input is not effort-neutral or time-neutral for the user, since both effort and time have been expended to no effect. Consequently, a score of -1 to reflect the wasted effort may be more appropriate. Similarly, a misrecognised input will most likely require a corrective action to either dismiss or undo the result of the incorrect input and an additional input action made to re-attempt the original desired input. Consequently, a score of -2 or -3 may be a more realistic reflection of the original input plus the corrective action plus the re-attempted input required.

3.5 Benchmarking the Useful Interaction Rate

If the notion of the bit rate of useful information transfer is taken as the most appropriate measure for benchmarking the practical acceptability of an assistive input system, then it is further possible to establish a baseline to compare the bit rate against.

As discussed earlier, the most basic working input system for almost all users with severe motor impairments is the simple binary switch used in conjunction with a scanning on-screen keyboard. Each successful binary switch input will generate 1 bit of information by definition. It is known from the work on the Model Human Processor [14] that for an able-bodied user the typical response time to a stimulus is ≈250ms, where the perceptual response time (τ_p) ≈100ms, cognitive cycle time (τ_c)

≈70ms and motor response time (τ_m) ≈70ms. Thus, if we assume no prediction, the idealized input interaction for an able-bodied user would look something like:

$$\text{Time\_per\_input} = \tau_p\ [\text{see the choice}] + 2\ \tau_c\ [\text{identify each of the options}] + \tau_c\ [\text{decide on which option}] + \tau_m\ [\text{operate the switch}] + f(t) \quad (9)$$

where f(t) is the mean time for the scanning input to land on the option to be selected. In the limiting case, and without the ability to predict ahead, the fastest scanning speed possible is anticipated to be 250ms per target. If standard able-bodied performance parameters are used in equation (9), the mean idealized time per bit of useful information using such a scanning keyboard is approximately 100ms + 140ms + 70ms + 70ms + 250ms = 630 ms (from equation 9), giving a useful information transfer bandwidth of (1/0.63) = 1.59 bits/s.

The values used above were derived for able-bodied users. The comparable values for motor impaired users have also been determined empirically [39] and are shown in Table 1.

Table 1. The Model Human Processor components and their observed values from Card, Moran and Newell [14] and Keates et al. [39] for able-bodied and motor impaired users.

Model Human Processor component	**Able-bodied (ms)**		**Motor impaired (ms)**
	[8]	**[31]**	**[31]**
Perception, τ_p	100 [50 – 200]	80 [70 – 100]	100 [70 – 120]
Cognition, τ_c	70 [25 – 170]	90 [90 – 100]	110 [100 – 130]
Motor function, τ_m	70 [30 – 100]	70 [60 – 80]	210 [100 – 310]
Simple reaction time ($\tau_p + \tau_c + 2\ \tau_m$)	310 (predicted)	310 (predicted) 320 (observed)	630 (predicted) 620 (observed)

As can be seen from Table 1, typical values for each of the Model Human Processor parameters were found to be: perceptual response time (τ_p) ≈100ms, cognitive cycle time (τ_c) ≈110ms and motor response time (τ_m) ≈110, 210 or 310ms, depending on the severity of the user's impairment, with increased severity leading to increased motor response times. From these values, it can be seen that a baseline idealized interaction time for the binary switch/scanning input is approximately 100ms + 220ms + 110ms

+ 110|210|310ms + f(t). Note that f(t) may have to be varied to allow for the range of reaction times, i.e. 320ms, 420ms or 520ms depending on the severity of the impairment and thus also the associated motor function time.

Consequently, using these assumptions, the best-case interaction rate for a user with a motor impairment is (1/0.86) = 1.16 bits/s (based on τ_m = 110ms). For users with severe motor impairments, that rate decreases to (1/1.05) = 0.95 bits/s.

For comparison, Table 2 shows the bit rates obtained from the gesture input system reported in [30].

Table 2. The useful interaction bit rate for four input modes for head and hand gesture recognition from [30]. Single Mode used only gesture from a single part of the body, either the user's hand or head and had a vocabulary of 6 gestures (left, right, up, down, yes, no). Duplicated Mode required the user to produce the same gesture with both their head and their hand, either concurrently or sequentially (e.g., first on the head and then on the hand) and had the same vocabulary of gestures. Different Mode required the user to produce Gesture 1 with their head and Gesture 2 with their hand. This mode increased the possible vocabulary size from 6 to 6*6 gestures, but also increased the cognitive and physical load on the user.

Input strategy	**Input vocabulary size**	**Useful interaction bit rate**
Single Mode - Head	6	0.72
Single Mode - Hand	6	0.77
Duplicated Mode – Head and Hand	6	0.65
Different Mode – Head and Hand	36	0.56

It can be seen from these calculations that the binary switch and scanning input outperforms the gesture input system where the bit rates seen ranged from 0.56 bits/s to 0.77 bits/s for the different types of input modes, combinations and vocabulary sizes used. As a recommendation, an interaction rate of 1 bit/s is suggested as the lowest baseline comparison. Any input system that fails to meet this target will struggle to claim acceptable performance.

3.6 Measuring the Effects of Cognitive Loading

Looking at Table 2, it can be seen that the Single Mode gestures had higher useful interaction bit rates than either Duplicate Mode or Different Mode. The Duplicate Mode was designed to improve the amount of useful data generated by facilitating the user to generate the same data at the same time on two channels, i.e. by the head and by the hand. The theory was that if one channel was not recognised, the second channel would provide the information necessary for the interaction to proceed. Consequently, the amount of useful information generated per unit time should have increased. The Different Mode was also designed to achieve an increase in the amount of useful interaction data generated by increasing the vocabulary size six-fold, from 6 gestures to 36 combined gestures, thus making every recognised pair of gestures (one on the head and one on the hand) convey more information per input.

However, both Duplicate and Different Modes produced lower useful interaction bit rates than the two Single Modes, despite being more data/information rich. The reason for this discrepancy was the increased cognitive load on the users. Observations of the users showed that while they were often perfectly happy to make the gestures in Single Mode, presenting more than one gesture at a time caused them significant difficulties. Those difficulties could even result in the users freezing completely as they were unable to translate the instructions into physical movements. Even trying to re-separate out the linked gestures into two separate gestures did not fully overcome this effect. So, for example, while the users were often content to make a Left gesture with their head, asking them to make a Left with both their head and hand at the same time caused difficulties. Some users, typically the less severely impaired, were able to try this, however some simply could not. In the latter case, the instructions were changed to be one gesture, e.g. head Left, and then the second one, hand Left, sequentially. The physical load was no more than two Single Mode inputs, though the users were still not able to reach the rates of useful information transfer that they achieved under the pure Single Mode entry. Consequently, it would appear that even just linking the notion of two separate gestures together adversely affects the interaction compared with presenting the same input as two wholly independent gestures.

A possible insight into what is happening is offered by the Model Human Processor study. As seen in Table 1, the motor function time for the users with motor impairments varied quite widely from that predicted by theory and from the values obtained for able-bodied users. Looking at the individual task times from one user, Figure 2, shows that the data is not spread uniformly in a normal-type distribution. Instead, the times for the button down and button up actions are a series of discrete peaks, rather than a typical bell curve.

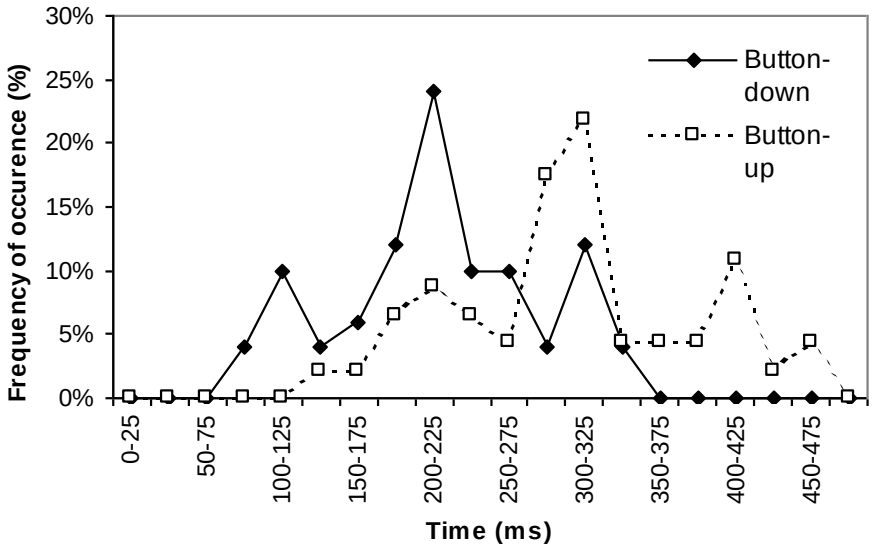

Figure 2. The motor response time for a user with a severe motor impairment showing button down and button up times for a repeated button press activity, specifically pressing and releasing a button 20 times as quickly as possible. The predicted time is c. 100ms, but distinct peaks can be seen at 100ms, 200ms, 300ms and 400 ms.

One possible interpretation of Figure 2 is that the assumption that an activity such as repeated key pressing is a purely automatic action and thus only has motor response time components, τ_m, present is incorrect. The peaks are distributed approximately multiples of cognitive cycle times, τ_c, apart. If this interpretation is correct, it means that users with severe impairments find it very difficult to achieve the fully automatic movements that much HCI theory expects. It also means that there is an elevated level of baseline cognitive load on the users from their motor impairments. A further step in the logic would then suggest that if a simple up-down action places cognitive demands on the users, a more complex action would place an even more elevated level of cognitive demand on the user. At this stage, it is difficult to know if this would be a linear increase in demand or a geometric or exponential increase. What is

clear, though, is that the Duplicate and Different Modes in the gesture input system had reached a tipping point in cognitive load for some of the users.

This is an area that needs further research. However, for the purposes of evaluating the acceptability of a new input system, designers and developers need to consider the cognitive loads placed on the users. Since this is difficult to measure absolutely, a comparative approach is the best option. It is suggested that the input system be evaluated with minimal and increased cognitive load on the user. One option might be to distract the user with another task, such as reciting a poem or recalling a list of instructions.

The same conditions should be evaluated with the benchmark input system. The reduction in user performance under the increased cognitive load condition, if any, for the new input system needs to be no worse than any reduction seen for the benchmark system under the same cognitive load.

3.7 Coping with User Variability

It is a common circumstance in most research in Universal Access that population sizes from any user evaluation sessions are likely to be quite small. Furthermore, user variability is also likely to be quite high. Apart from the expected between-users variability, individual users themselves can exhibit notable changes in their capability over time and may, for example, require medical treatment in the course of an extended set of user evaluation sessions. They may become fatigued easily and sometimes cannot complete trials or experimental conditions. Only users who are already used to interacting with computers may be suitable if time is restricted. Consequently, it is usually necessary for experimenters to run any trials on a long-term basis, develop a working relationship with the users and to keep experimental conditions as constant as possible. Repeated measures designs should generally be employed. Obviously, these practical difficulties can give rise to missing data problems resulting from incomplete conditions, causing the loss of levels and factors from designs, and making the systematic varying of conditions in empirical studies difficult. In addition, the increased range and skewed variability resulting from the

range of functional impairments, can lead to increased noise and violation of the standard assumptions of statistical tests.

Any attempt at empirical evaluation must be sufficiently robust to cope with both of these factors, which would otherwise limit the usefulness and applicability of detailed statistical analyses. Again, one of the strongest recommendations that can be made here is to allow the users as much time as possible using the system to be evaluated to at least limit the effects of the process of learning [14].

Where statistical tests are possible without violation of standard assumptions, such as normality of distribution or homogeneity of variance, they should be carried out. However, the statistical power of these experiments may be highly variable because of the reasons outlined and the small sample size. Despite the inherent variability, though, effect sizes can often be large [e.g. 39]. For this reason, some statistical results that may not be significant at this level can be analysed in terms of statistical power (1 - β: the probability of rejecting a false null hypothesis), and estimates of effect size given [40].

4 A Suggested Set of Metrics

It is difficult to suggest a comprehensive universal methodology for assessing the likely acceptability of a new input system. However, it is certainly possible to suggest best practice guidance. Table 3 shows suggested metrics to consider.

Table 3. The factors and metrics to consider when evaluating the overall acceptability of a new input system for users with functional impairments.
* Note – effects of cognitive and physical loading – in other words, does user performance vary with the system when the user is subject to different cognitive and physical load conditions, e.g. does asking them to remember something or to undertake a concurrent activity while performing the task affect their performance more than might usually be expected?

Attribute to consider	**Metrics/factors to consider**
Practical acceptability	• Time to complete an action

	• Success rate in completing an action • Time to complete a task • Success rate in completing a task • Frequency and nature of errors • Severity of consequences of errors • Throughput • (Useful) Information transfer rate • Fatigue effects and rate of fatigue • Effects of cognitive and/or physical loading* • Potential for RSI and other injuries
Social acceptability	General satisfaction measures (e.g. TLX), but also benchmarking against any competitors, e.g.: • Which did you consider easier to use? • In what ways was the new system better than your existing solution? • In what ways was it worse? • Which would choose to use? Why? • What would you change about the new system?
Overall acceptability	• Is the input system "good enough"? • Are there positive outcomes for practical and social acceptability? • Are those positive outcomes truly representative of genuine use or are they only for specific experimental conditions?

In terms of using Table 3 to determine whether a new input system is likely to be considered acceptable, it is suggested that the interaction rate, calculated as in section 3.4 above, is a good indicator of whether a system has the potential to be successful. If error rates are considered to be a major factor, then a modified interaction rate using a modified "score" to reflect the full impact of the errors, of the type shown in equation (8), should be used. This modified bit rate can be thought of as the "useful" interaction bit rate, i.e. the bit rate that is actually moving the interaction forward towards its completion. The useful interaction rate generated by the new input system

should be benchmarked against any competitor system. In the case where the users are not able to use competitor systems, for example where they have very severe motor impairments, the very lowest interaction rate that should be considered acceptable is 1 bit/s, i.e. the rate achieved by a binary switch and scanning input system.

For the social acceptability, again benchmarking is key. Social acceptability is typically established through questionnaires, surveys and interviews where researchers try to establish user preferences, likes and dislikes [e.g. 41]. For a new system to prove successful, it needs to at least match or ideally outperform its competitors by scoring more highly on the user preferences. However, it can be difficult to dislodge a long established and dominant input system from a user's preference. There is an in-built tendency within users to prefer that which is familiar [e.g. 42]. This tendency appears to be especially true for skills that require substantial periods and efforts to learn. In the case of a user with a functional impairment, it is reasonable to assume that the acquisition of skills to master any input system could have been significant and the desire to learn and/or adapt to a new system is correspondingly decreased.

For example, consider alternative input actions on a touchscreen, where the entire interface is built around the notion of the user simply tapping on the icon or button they wish to select and/or activate. It is unrealistic, though, to expect the users to unlearn that dominant form of interaction with a touchscreen in favour of an innovative approach, such as lift-off (where it is the point where the finger leaves the screen that is activated, not where it first touches) or circling the target, within a 30 to 60 minute user trial session. Consequently, the perceived social acceptability of the new system, as measured through the usual approaches such as Likert scales [43], may not be as high as might otherwise be expected as users could still prefer their existing input technologies that may have been previously learned at great effort. The only solution to this issue is likely to be significantly extended periods of use with the new system, to allow users to acquire comparable familiar and competence to their more usual input system. In other words, it may be necessary to evaluate social acceptability once the full learning process of the new technology is complete and not too early in the evaluation process. Research, for example, in Japan has shown a

positive correlation to user acceptance of new technologies and frequency/duration of use for older adults [44].

Otherwise, the new technology is potentially being unfairly penalised because of the in-built preference towards the familiar rather than the new. Only technologies that offer something profoundly improved or new can overcome this bias on a shortened timescale and such technological leaps in input technology are rare (for example, the Microsoft Kinect system does away for the need for the user to hold an input device). However, research has shown that even with such radical advancements, users still often prefer more traditional input methods [e.g. 45].

Hence, to gain more truly representative estimates of both social and practical acceptability, it is strongly recommended that users are given time to fully adjust to any new input system. This can either be accomplished by giving them access to the new input system prior to the user trial session or by holding multiple sessions over a number of days. Figure 1 shows the improvement in performance over repeated trials than can usually be expected and follows the type of improvement predicted by the Power Law of Practice [14]. Most research papers in universal access seem to typically report results from one-off user trial sessions.

5 Conclusions

To improve the success of assistive input systems outside of the research laboratory, it is necessary for researchers and developers to take a more sophisticated view of how well the systems that they are developing genuinely meet the needs of the users. While methods for assessing the social acceptability of such systems are widely understood, although not necessarily undertaken, there is much more variability over the approaches to measure the practical acceptability of such systems.

This paper has discussed the notion of other measures, such as the bit rate of useful information transfer, as more sophisticated metrics than the recognition rate typically reported in many papers on universal access. It has also introduced a method for establishing a straightforward baseline for such a measure to be compared with.

Overall, the use of such more complete metrics, such as those listed in Table 3, would help designers and researchers understand the likely success or otherwise of a new assistive input system more clearly than the metrics that currently prevail.

References

1. Polacek O, Sporka AJ, Slavik P (2015) Text input for motor-impaired people. Int J on Universal Access in the Information Society. (Springer online first)
2. Nicolau H, Guerreiro T, Lucas D, Jorge J (2014) Mobile text-entry and visual demands: reusing and optimizing current solutions. Int J on Universal Access in the Information Society, 13(3), pp 291-301
3. Paradise J, Trewin S, Keates S (2005) Using pointing devices: Difficulties encountered and strategies employed. Proceedings of the 3rd International Conference on Universal Access in Human-Computer Interaction, (Las Vegas, NV July 2005), pp 22-27
4. Keates S, Trewin S (2005) Effects of Age and Parkinson's Disease on cursor positioning using a mouse. Proceedings of 7th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2005), Baltimore, MD, October 2005, pp 68-75
5. Stephanidis C, Savidis A (2001) Universal Access in the Information Society: Methods, Tools, and Interaction Technologies. Int J on Universal Access in the Information Society, 1(1), pp 40-55
6. Begnum MEN, Begnum K (2012) On the usefulness of off-the-shelf computer peripherals for people with Parkinson's Disease. Int J on Universal Access in the Information Society, 11(4), pp 347-357
7. Rodrigues E, Carreira M, Goncalves D (2014) Enhancing typing performance of older adults on tablets. Int J on Universal Access in the Information Society, Springer, (Springer Online First) DOI: 10.1007/s10209-014-0394-8
8. Keates S, Potter R, Perricos C, Robinson P (1997) Gesture recognition - Research and clinical perspectives. Proceedings of RESNA 97, Pittsburgh, PA, pp 333-335
9. Nielsen J (1993) Usability engineering. Morgan Kaufman, San Francisco
10. Mourouzis A, Antona M, Stephandis C (2011) A diversity-sensitive evaluation method. Universal Access in the Information Society, 10(3), pp 337-356
11. Keates S (2007) Designing for accessibility: A business guide to countering design exclusion. CRC Press, Boca Raton, FL
12. Buhler C (1998) Robotics for rehabilitation - a European (?) perspective. Robotica, 16(5), pp 487-490
13. Tsourakis N (2014) Using hand gestures to control mobile spoken dialogue systems. Int J on Universal Access in the Information Society, Springer, 13(3), pp 257-275
14. Card SK, Moran TP, Newell A. (1983) The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Hillsdale, New Jersey

15. Keates S, Clarkson PJ (2003) Countering design exclusion: Bridging the gap between usability and accessibility. Int J on Universal Access in the Information Society, Springer, 2(3), 215-225
16. Freedom Scientific (2014) JAWS for Windows: Screen reading software. Available at: http://www.freedomscientific.com/Products/Blindness/JAWS
17. Keates S, Adams R, Bodine C, Czaja S, Gordon W, Gregor P, Hacker E, Hanson V, Kemp J, Laff M, Lewis C, Pieper M, Richards J, Rose D, Savidis A, Schultz G, Snayd P, Trewin S, Varker P (2007) Cognitive and learning difficulties and how they affect access to IT systems. Int J on Universal Access in the Information Society, Springer, 5(4), 329-339
18. Gregor P, Dickinson A (2006) Cognitive difficulties and access to information systems: an interaction design perspective. Int J on Universal Access in the Information Society, 5(4), pp 393-400
19. El Kaliouby, Robinson P, Keates S (2003) Temporal context and the recognition of emotion from facial expression. Proceedings of HCI International 2003, Crete, Greece, pp 631-635
20. Keates S, Varker P, Spowart F (2011) Human-machine design considerations in advanced machine-learning systems. IEEE/IBM Journal of Research and Development, IEEE, 55(5) September/October 2011, pp 4:1-4:10
21. Welton T, Brown DJ, Evett L, Sherkat N (2015) A brain-computer interface for the Dasher alternative text entry system. Int J on Universal Access in the Information Society, Springer, (Springer online first)
22. Hwang F, Keates S, Langdon P, Clarkson PJ (2005) Movement time for motion-impaired users assisted by force-feedback: effects of movement amplitude, target width, and gravity well width. Int J on Universal Access in the Information Society, Springer, 4(2), pp 85-95
23. Trewin S, Keates S, Moffatt K (2008) Individual responses to the Steady Clicks cursor assistance technique. Disability and Rehabilitation: Assistive Technology, Informa Healthcare, 3 (1 & 2) January 2008, pp 2-21
24. Sandnes FE, Tan TB, Johansen A, Sulic E, Vesterhus E, Iversen ER (2012) Making touch-based kiosks accessible to blind users through simple gestures. Int J on Universal Access in the Information Society, 11(4), pp 421-431
25. Keates S (2015) A pedagogical example of teaching Universal Access. Int J on Universal Access in the Information Society, Springer, 14(1), pp 97-110
26. International Standard Organization (ISO) (1998) ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs), Part 11: Guidance on Usability Specification and Measures. Technical report. ISO, Geneva
27. Frokjaer E, Hertzum M, Hornbaek K (2000) Measuring usability: Are effectiveness, efficiency and satisfaction really correlated? Proceedings of CHI 2000, The Hague, NL, pp 345-352
28. Keates S, Clarkson PJ (2003) Countering design exclusion: an introduction to inclusive design. Springer-Verlag: London
29. Jastrzembski T, Charness N, Holley P, Feddon J (2005) Input devices for web browsing: age and hand effects. Int J on Universal Access in the Information Society, 4(1), pp 39-45

30. Trewin S, Pain H (1999) Keyboard and mouse errors due to motor disabilities. Int J of Human-Computer Studies, 50(2), pp 109-144
31. Hu R, Feng J, Lazar J, Kumin L (2011) Investigating input technologies for children and young adults with Down syndrome. Int J on Universal Access in the Information Society, 12(1), pp 89-104
32. MacKenzie IS, Soukeroff RW (2002) A character level error analysis technique for evaluating text entry methods. Proceedings of the Second Nordic Conference on Human-Computer Interaction (NordiCHI '02), ACM, pp 243-246
33. Soukoreff RW, MacKenzie IS (2003). Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2003), Fort Lauderdale, FL, pp 113-120
34. Hwang F, Keates S, Langdon P, Clarkson PJ (2005) A submovement analysis of cursor trajectories. Behavior and Information Technology (BIT), Taylor & Francis, 24(3), pp 205-217
35. MacKenzie IS, Kauppinen T, Silfverberg M (2001) Accuracy measures for evaluating computer pointing devices. Proceedings of CHI 2001, Seattle, WA, pp 9-15
36. Keates S, Hwang F, Langdon P, Clarkson PJ, Robinson P (2002) The use of cursor measures for motion-impaired computer users. Int J on Universal Access in the Information Society (UAIS), Springer, 2(1), November, 2002, pp 18-29
37. Holbert B, Huber M (2011) Design and evaluation of haptic effects for use in a computer desktop for the physically disabled. Int J on Universal Access in the Information Society, 10(2), pp 165-178
38. Keates S, Robinson P (1999) Gestures and multimodal input. Behaviour and Information Technology, Taylor and Francis Ltd. January-February, 1999. 18(1), pp 36-44
39. Keates S, Langdon P, Clarkson PJ, Robinson P (2002) User models and user physical capability. User Modeling and User-Adapted Interaction (UMUAI), Wolters Kluwer Publishers 12(2-3), pp 139-169
40. Chin D (2001) Empirical Evaluation of User Models and User-Adapted Systems. User Modeling and User-Adapted Interaction, 11(1-2), pp 181-194
41. Murphy E , Kuber R, McAllister G, Strain P, Yu W (2008) An empirical investigation into the difficulties experienced by visually impaired Internet users. Int J on Universal Access in the Information Society, 7(1), pp 79-91
42. Palacio RR, Acosta CO, Cortez J, Morán AL (2015) Usability perception of different video game devices in elderly users. Int J on Universal Access in the Information Society. (Springer online first)
43. Melo P, Jorge L (2014) Quantitative support for UX methods identification: how can multiple criteria decision making help? Int J on Universal Access in the Information Society, 14(2), pp 215-229
44. Umemuro H, Shirokane Y (2003) Elderly Japanese computer users: assessing changes in usage, attitude, and skill transfer over a one-year period. Int J on Universal Access in the Information Society, 2(4), pp 305-314

45. Heloir A, Nunnari F (2015) Toward an intuitive sign language animation authoring system for the deaf. Int J on Universal Access in the Information Society. (Springer online first)