Applied Ergonomics 45 (2014) 886–894

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

A comparison of methods currently used in inclusive design

Joy Goodman-Deane [a,*], James Ward [a], Ian Hosking [a], P. John Clarkson [a]

[a] Engineering Design Centre, Engineering Department, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK

ARTICLE INFO

Article history:
Received 11 July 2012
Accepted 11 November 2013

Keywords:
Inclusive design
Methods
Exclusion

ABSTRACT

Inclusive design has unique challenges because it aims to improve usability for a wide range of users. This typically includes people with lower levels of ability, as well as mainstream users. This paper examines the effectiveness of two methods that are used in inclusive design: user trials and exclusion calculations (an inclusive design inspection method). A study examined three autoinjectors using both methods ($n = 30$ for the user trials). The usability issues identified by each method are compared and the effectiveness of the methods is discussed. The study found that each method identified different kinds of issues, all of which are important for inclusive design. We therefore conclude that a combination of methods should be used in inclusive design rather than relying on a single method. Recommendations are also given for how the individual methods can be used more effectively in this context.

1. Introduction

Inclusive design focuses on making mainstream products and services usable by as many people as is reasonably possible, without requiring them to use specialised adaptations (Keates and Clarkson, 2003). It thus seeks to meet the needs of a wide range of users, including both mainstream users and those with specific needs. In particular, it typically aims to include more of those with lower levels of sensory, motor and cognitive ability.

A range of methods are commonly used in inclusive design, including both general user-centred design methods and methods specifically developed for inclusive design. In particular, user trials are widely considered to be one of the most reliable methods for identifying usability problems, both within user-centred design as a whole (e.g. Nielsen and Landauer, 1993; Ebling and John, 2000) and within inclusive design in particular (e.g. Cardoso et al., 2005). These trials are commonly used in design practice with limited samples (Goodman-Deane et al., 2010a,b). For example, Sims (2003) found that many designers "only involve a few users, who may not reflect the variety of needs of the target user group". If more representative samples are sought, it seems likely that practitioners will follow standard practice in mainstream design by looking for a good spread of demographic variables such as age and gender.

Some have expressed concern about the effectiveness of this approach within inclusive design. For example, Grudin (2006, p. 662) notes that "a [design] team relying on usability studies, for example, is unlikely to inquire very deeply into the diversity of the participants, few or none of whom they might ever see". Further work is needed to determine whether user trials, used in this fashion, are adequate to uncover the main usability problems in an inclusive design context.

Some specialised inclusive design methods have also been developed. In this paper, we focus on exclusion calculations as an example of an expert appraisal method developed for use in inclusive design (Keates and Clarkson, 2003; Waller et al., 2010). This method helps usability experts to assess inclusivity. The calculations estimate how many people in the target population would be excluded from using a product or service due to limited user capabilities. This process is sometimes referred to as an "exclusion audit", although this term can also refer to a package of methods, including user trials and expert appraisal as well as exclusion calculations.

Exclusion calculations have been used successfully in both research and commercial contexts, along with other methods, such as user trials (e.g. Klein et al., 2003; Clarkson et al., 2007). These studies combined findings from the different methods to build overall pictures of the products' usability. In fact, Clarkson et al. (2007) recommend that exclusion calculations be used along with user trials and expert appraisal to provide the different kinds of guidance needed for inclusive design. However, this recommendation has not been tested extensively: there has only been limited work comparing the results of user trials and exclusion calculations. Furthermore, there has been limited work in exploring how effective exclusion calculations actually are for inclusive design, particularly at identifying usability issues.

For example, Cardoso (2005) (c.f. Cardoso et al., 2005) examined the usability of electric kettles, domestic heating controls and

* Corresponding author.
E-mail address: jag76@cam.ac.uk (J. Goodman-Deane).

0003-6870
http://dx.doi.org/10.1016/j.apergo.2013.11.005

digital television boxes. He compared the usability problems identified by exclusion calculations and user trials. However, he used the user trial results as a benchmark to identify the main usability problems, rather than fully comparing the contributions of the different methods. He also used earlier versions of the exclusion calculations.

More recently, Combe et al. (2012) compared the results of exclusion calculations and user trials in assessing the usability of digital programmable thermostats. However, they focused on the exclusion estimated by the calculations rather than the usability problems found, and thus did not examine the full use of this method. This also made it difficult to compare the methods, as they did not use the same outcome measures from the two methods.

This paper aims to address these gaps in the research by examining the effectiveness of both user trials and exclusion calculations more extensively. These methods use different measures, such as timings and exclusion figures. However, they are both commonly used for the same ends, particularly to uncover usability and accessibility problems. We therefore compared the methods on their ability to achieve these goals. A study was conducted involving both methods, and the usability problems and other findings identified by each method are compared. The effectiveness of the methods for inclusive design is also discussed.

2. Study aims and methods

The primary aim of the study was to compare the usability of three autoinjectors, particularly in the context of inclusive design. However, it also provided an opportunity to compare different inclusive design methods in practice, and this paper focuses on this latter aim. Full details of the study can be found in an internal technical report (Goodman-Deane et al., 2010a,b).

One limitation of the study is that the exclusion calculation method was developed by some of the authors. We have tried to address this by describing the methods, conclusions and reasoning in detail, but care should still be taken in interpreting the results.

In addition, the study was commissioned and funded by Oval Medical Technologies, who designed one of the autoinjectors. However, this paper does not focus on which device was the most inclusive, but on the effectiveness of the methods used, which should be independent of this conflict of interest. Furthermore, the study was commissioned as an independent assessment and thus care was taken to be impartial, e.g. by presenting the devices in identical boxes during the user trials. The neutrality of the investigation was explained to the company, and it was agreed that the results would be published regardless of the findings.

2.1. Devices examined

An autoinjector is a medical device which delivers medicine through the skin using a needle. The autoinjectors examined were home-use devices designed for patients to inject themselves, typically once every few weeks. Three devices were compared (see Fig. 1):

- Device A: a trainer version of the HUMIRA® autoinjector (Abbott Laboratories, 2013);
- Device B: a trainer version of the ENBREL SureClick autoinjector (Amgen Inc., 2013);
- Device C: a prototype of an autoinjector designed by Oval Medical Technologies (Oval Medical Technologies, 2013).

Devices A and B were chosen as leading market representatives of this type of autoinjector. Device C was a new design produced by Oval Medical Technologies.

2.2. Expert appraisal

Firstly, an expert appraisal was conducted, involving a task analysis (Kirwan and Ainsworth, 1992) and qualitative risk analysis for the entire lifecycle of each device. This involved the authors working carefully through the steps involved in using the devices, considering which steps were most likely to fail, how failure could occur, and what effects failure might produce. The results informed the tasks examined in the subsequent user trials and exclusion calculations, and the observations taken in the user trials (see the following sections).

The full analysis cannot be shown for reasons of space, but a summary of the core stages of device use can be found in Table 3 in Section 3.2.3. In summary, the user removes a device from its packaging, checks it, and prepares the injection site. He or she then removes a cap or caps from the device, presses the device against the skin, presses a button (except for device C which activates automatically), and holds the device in place until the injection is complete, before tidying up. These were the stages examined in the user trials and exclusion calculations because they were the core activities related to the autoinjectors themselves.

The task analysis reflects the devices' recommended use. In practice, some steps can be done in a different order. For consistency, some assumptions were made about how the medicine is delivered and stored, and how patients remind themselves to take the medicine.

2.3. User trials

2.3.1. Sample

User trials are commonly used in design practice with limited samples, but this study required a sample more representative of the population as whole. To obtain this, a quota sampling strategy was used, aiming for a distribution of genders, ages and education levels close to that in England as a whole. As a result, these user trials may be more comprehensive than those typically conducted in design practice. Nonetheless, they represent what many designers may consider good practice in user trials.

Of the 30 participants, 13 were male, and the age distribution was: 18–39 ($n = 8$), 40–59 ($n = 11$), 60+ ($n = 11$). The education levels were: degree or equivalent ($n = 14$), 2 A-levels or equivalent ($n = 5$), 5 GCSEs or equivalent ($n = 6$), fewer qualifications ($n = 5$). Only one participant had any prior experience of autoinjectors and this experience was very limited. Participants were paid £15.

Fig. 1. The autoinjectors used in the study: trainer versions of the HUMIRA and ENBREL SureClick autoinjectors, and a prototype of a new autoinjector.

2.3.2. Method

Each trial was conducted by the same facilitator, working from a script to ensure consistency. Two observers in another room took notes from a video and audio feed.

After giving consent, participants used each of the autoinjectors in turn. The order of the autoinjectors was counterbalanced. For consistency and practicality, the injection was performed on an injection trainer attached to the participants' thighs over their clothing (Fig. 2), rather than on their skin. An injection trainer is a piece of fake skin and muscle designed for practising giving injections.

For each autoinjector, participants were given a white box marked A, B or C depending on the device inside. White boxes were used to reduce irrelevant differences between devices and because a production-quality box for device C was not available. Because the peel-off strip on the plastic tray for device A could not be replicated realistically, the devices were provided without trays. The boxes also contained an alcohol wipe and instructions for using the device (see Fig. 3).

The instructions were produced by Oval Medical Technologies by extracting verbatim the section(s) on device use from the patient information leaflets for devices A and B. A similar section was written by Oval for device C. The instructions were printed on 50gsm A4 paper at around 8 point, for similarity with real information leaflets. These abbreviated leaflets were used to reduce the amount of reading required and differences in the layout and quantity of information.

Participants were asked to imagine that the box was already warmed to room temperature, and to follow the instructions in the box to perform an injection on the injection trainer. They followed the steps in Table 3 up to disposing of the waste, and so were also provided with a sharps bin and cotton wool.

After using each device, participants completed a NASA Task Load Index (TLX) questionnaire (Hart and Staveland, 1988), as a measure of perceived workload. This involved rating their experience of using the device on six scales: Mental demand, Physical demand, Temporal demand, Perceived performance, Effort and Frustration.

At the end of the trial, they were asked which autoinjector they preferred and why, and which they found easiest and most difficult to use. They also provided some information on demographics and capabilities (discussed in Section 4.1.1).

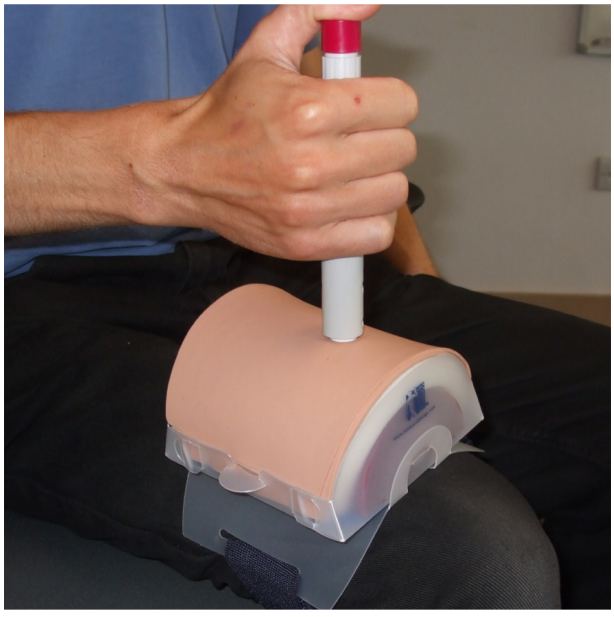

Fig. 2. Injection trainer.

2.3.3. Observations

One observer noted times of key events during device use, which were identified in advance by the expert appraisal (Section 2.2). Both observers also took notes on participants' behaviour, especially errors and unusual actions in device use. They particularly looked out for the errors identified in the expert appraisal. The facilitator also noted some errors. Observers and facilitator compared notes after each session to identify the main errors in device use. Where there was discrepancy in timings or observations, the video was re-examined to clarify what had actually happened.

2.4. Exclusion calculations

An exclusion calculation was performed by the authors for each autoinjector. A particular aim was to examine issues for less able users. The method is based on the principle that people are excluded from using a product if their capabilities are less than those demanded by the product, given the environmental context (Persad et al., 2007). For example, a mobile telephone may require a certain level of dexterity. Someone with low dexterity capability would be unable to press its buttons accurately, and thus be effectively excluded from using it.

A calculation involves examining the demand a product places on various user capabilities. It then estimates the proportion of the target population whose capabilities do not meet these demands and thus would be unable to use the product (Keates and Clarkson, 2003; Waller et al., 2009).

2.4.1. Procedure

The exclusion calculations examined the core stages of device use (see Table 3). These were the same activities examined in the user trials, with the addition of removing the device from its tray and checking for side effects, as they were not prevented by the same practical concerns.

Each task was examined separately. The authors rated the demand placed by that task on 29 user capabilities in five categories: vision, hearing, thinking, dexterity & reach, and mobility. The capabilities correspond to data from the Disability Follow-up Survey (DFS) (Grundy et al., 1999). The detailed capabilities can be found in (Engineering Design Centre, 2012). Abbreviations are also given in Table 2 (Section 3.2.1).

The thinking capabilities relate to the ability to perform specific cognitive tasks, such as "count well enough to handle money". The demands on these capabilities were rated 0 or 1 depending on whether the capability is needed to perform the product task. The demands on the other capabilities were rated 0 (low), 1, 2 or 3 (high), depending on the level of that capability required by the task. The points 0, 1, 2 and 3 were defined based on data from the DFS.

These demands were then compared with the capabilities of people in the DFS database. An algorithm identified those people in the database whose capabilities do not meet the product's demands and thus would be unable to use the product. These people were added up to give a total number excluded. The database is representative of the 1996/97 British adult populations living in private households, and so this number could be scaled up (with appropriate weighting) to estimate the percentage of this target population who would be unable to use the product (Waller et al., 2010). More details on exclusion calculations can be found in (Waller et al., 2009, 2010).

Fig. 3. The devices and related materials given to participants in the user trials.

3. Results

3.1. User trial results

3.1.1. Times

Various times were calculated for each user. Means are shown in Fig. 4. All times except activation time (the time for which the device was on the skin after activating it) differed significantly between the devices ($p < 0.001$, one-factor ANOVAs). It should be noted that times are less useful in inclusive design than mainstream design, as people with limited capabilities often take longer to complete tasks (see Section 4.4). Nevertheless, times are commonly used, and can still provide a comparison between different devices for the same person and so were included in this study.

3.1.2. Errors

Errors were categorised by the facilitator as follows.

- *Critical errors with 10s threshold*. The device failed to be in contact with the skin for 10 s after activation, contrary to the instructions. Various reasons for this are shown in Table 1. As a result, the user may fail to receive any medicine at all, or may receive the wrong amount.
- *Critical errors with 4s threshold*. As above, but with the time threshold reduced to 4 seconds. This is of interest because the time quoted in the instructions errs on the safe side. For example, device C can deliver the full dose in approximately 4 s. The time needed for the other devices is not known.

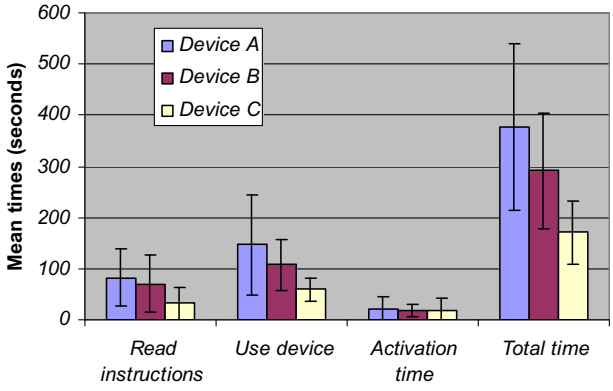

Fig. 4. Mean times involved in using the devices. Error bars indicate one standard deviation.

- *Minor errors and instances of confusion and difficulty*. Less critical errors. The user would receive the proper dose but may experience other problems.

Observing these errors can identify issues with the devices, although it should be noted that it is difficult to determine the exact causes of some of the problems. In particular, it can be hard to determine if errors were caused by the design of the instruction sheet or the physical design of the autoinjector, although these are not unrelated (see Section 4.1.2).

The critical errors are shown in Table 1. There is no significant difference between devices in the number of critical errors with a 10 s threshold ($p > 0.05$, Chi-squared) and a slightly significant difference with a 4 s threshold ($p < 0.05$, Chi-squared).

Too short activation times were common across all devices, especially with a 10 s threshold. Other critical errors varied between devices. In particular, a common error with device A was activating the device when not on the skin. This included activating it upside-down or with the cap still on. Seven people failed to activate device B at all, even though five of them had the device in the right position and applied significant pressure to the button. It seems that this device does not activate if it is not pressed hard enough to the skin prior to pressing the button.

3.1.3. Minor errors

Effort focused on capturing critical errors, so some minor errors may have been missed. Thus the numbers of these errors are not as informative as their nature.

About a third of the participants for each device had problems pinching or stretching the skin, e.g. omitting to pinch it, pinching it instead of stretching it or releasing the grip too soon. The skin on the

Table 1
Critical errors in using the autoinjectors.

Critical errors	Number of participants experiencing this difficulty with device		
	A	B	C
Failure to activate device	1	7	0
Activated device when not on skin	4	0	1
Too short activation time (<4 s)	1	3	1
Too short activation time (4–10 s)	4	2	4
Other	0	1	1
Total critical errors with 10 s threshold	10	13	7
Total critical errors with 4 s threshold	6	11	3

injection trainer was already taut, and so the devices still activated successfully. In real life, these errors would likely result in the devices failing to activate. It is unclear how users would respond to this in practice: some may correct themselves, while others may give up.

Another common error was twisting the device during the injection. In real life this would cause pain and disrupt the dose. This behaviour was observed with all three devices, but most often with device A. This may be because participants were trying to see the inspection window before they removed the device, as stated in the instructions for device A (see Table 3). Others just checked the window after removing the device. This may be partly due to confusion with the instructions for devices B and C, which said to check the window afterwards.

Some participants (with all devices) omitted checking the expiry date or drug quality, which could lead to receiving poor quality medicine. Others attempted to recap the device. This carries a risk of stabbing oneself with the needle.

Some people experienced significant confusion with the instructions, especially for device B. With device A, removing the caps often caused difficulty or confusion (9 participants), e.g. some participants removed a cap and then tried to remove the underlying button as it was also a cap.

3.1.4. Workload

Mean TLX scores are shown in Fig. 5. These measure perceived workload, with low scores indicating low workload. Overall unweighted or Raw TLX (RTLX) scores were calculated, as these have been shown to be an effective measure of overall workload (Byers et al., 1989). All scores differ significantly over the three devices ($p < 0.01$, one-factor ANOVAs).

3.1.5. Preferences

Twenty-nine out of 30 participants preferred device C, while one chose device A. The main reasons given were that device C was easy to use, had simple and clear instructions, was small/light-weight, and had fewer aspects to think about. Twenty-nine participants rated device C as the easiest to use, for similar reasons.

Twenty participants considered device A the hardest to use, saying that the instructions were long and complicated; and there were too many tasks to perform, particularly having two caps to remove. It was also hard to see the inspection window. Ten participants chose device B. Reasons included confusing instructions and comments on various aspects of the device.

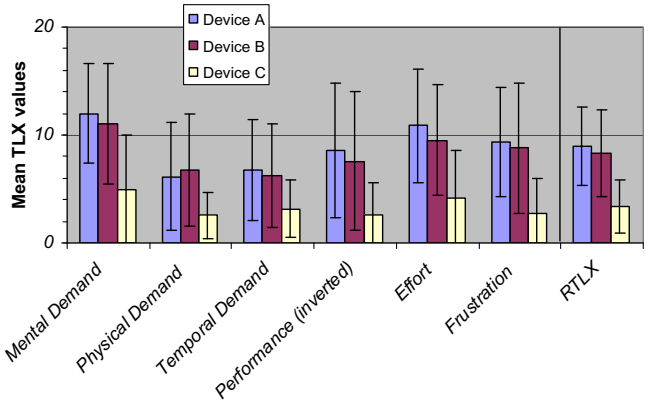

Fig. 5. Mean TLX scores (out of 20). Values for performance have been inverted so that low numbers indicate good performance (low workload). Error bars indicate one standard deviation.

3.2. Exclusion calculation results

3.2.1. Individual demand levels

Table 2 gives an example of the demand involved in an individual task. The full set of demand levels cannot be shown due to space constraints. In many cases, the demand values are the same for all the devices. Sometimes this is because the task is performed in a very similar (or identical) way, e.g. cleaning the injection site. In other cases, the demand levels are similar even though the task varies. This may be because the difference in demand is too small to be picked up on the capability scales available. For example, the cognitive demand in reading the instructions varies depending on the instructions. However, the scales only pick up the largest cognitive issues, so this task is rated the same for all the devices.

The main differences in individual task demands between the devices are as follows.

- The inspection window in device C is larger than in the other devices, so inspecting the drug quality is easier and the device does not have to be lifted up to the light.
- There is no need to press a button on device C, resulting in decreased dexterity demand.
- Device A has higher dexterity demands for waiting for the drug to be delivered, because the button should be kept depressed in addition to pinching the skin with one hand and holding the device in place with the other. At the same time, the patient should check the window on the device to see if the injection is complete.

3.2.2. Overall demand levels

Table 2 also gives the overall levels of demand. The calculations assume a linear, independent sequence of tasks. Thus the overall demand for a capability is the maximum of all the individual demands on that capability.

Some of the overall demands for the whole sequence are higher than those involved in just the device use. This is primarily due to the dexterity demands in cleaning the injection site and disposing of the waste, and the cognitive demands in reading the instructions.

3.2.3. Exclusion

The proportion of the population excluded by each of the tasks is shown in Table 3. The overall exclusion was calculated from the set of overall demands. Percentages are of the British (GB) adult population in 1996/97, based on the DFS (Grundy et al., 1999).

3.2.4. Discussion

Despite some differences in individual values, the overall demands and exclusion values for devices A and B are identical (see Tables 2 and 3). This is because the point of greatest demand on each capability is the same. For example, device B has a lower dexterity requirement than A for waiting until the drug is delivered, but this is masked by the same (higher) requirement for activating the device.

Device C does reduce exclusion slightly (by 0.6%). This is primarily due to lower dexterity demands on two of the most demanding tasks: checking drug quality and activating the device. However, these are partially masked by the demands in other tasks, especially cleaning the injection site, disposing of the packaging and reading the instructions. To get the full benefit from the design features of device C, the demands on these other tasks need to be reduced. If this were done, then the overall exclusion of device C would become 10.4%. Improving these other tasks would also reduce the exclusion for devices A and B, although only to 11.5% because it is limited by the demands of activating the devices.

Table 2
Demand levels for an example task and overall. 0 represents low, and 3 represents high demand. Blank entries are used where the demand is zero for all three devices. Names of capabilities are abbreviated. Values that differ across devices are in bold.

	Vis		Hear		Thinking												Dexterity & Reach									Mobility		
Tasks	Read	Recognise	Speech	Sound	Remember names	Remember turn off	Understand people	Express oneself	Write letter	Read article	Remember message	Watch TV	Hold Conversation	Count	Tell time	Keep focused	Think clearly	One hand pick up	Both hands pick up	One hand fine finger	Both hands fine finger	One arm briefly	One arm long time	Both arms briefly	Both arms long time	Bending	Walking	Steps	Balance
Example task: Activate device (while holding it in position and pinching skin)																													
Device A		**1**		3												1	1	1		**3**	**1**								
Device B		**1**		3												1	1	1		**3**	**1**								
Device C		**0**		3												1	1	1		**1**	**0**								
Overall demand (just device use)																													
Device A	3	1		3		1								1	1	1	1	1	1	**3**	1	**2**		1					
Device B	3	1		3		1								1	1	1	1	1	1	**3**	1	**2**		1					
Device C	3	1		3		1								1	1	1	1	1	1	**1**	1	**1**		1					
Overall demand (all tasks, including device-independent steps such as cleaning the injection site)																													
Device A	3	1		3		1			1	1			1	1	1	1	1	1	1	**3**	2	**2**		1					
Device B	3	1		3		1			1	1			1	1	1	1	1	1	1	**3**	2	**2**		1					
Device C	3	1		3		1			1	1			1	1	1	1	1	1	1	**2**	2	**1**		1					

The exclusion could be further reduced (to 8% for device C) by reducing the vision demand of several of the tasks, and reducing the demand of hearing the device activate, e.g. by making the activation click louder or providing other feedback.

Table 3
Exclusion values for each of the tasks. Values that differ between devices are in bold. Actual exclusion may be higher in practice, as the calculations do not identify all difficulties for people with no or mild impairments (see Section 4.2).

Tasks	% of total adult British population excluded		
	Device A	Device B	Device C
Take device from box	**6.6**	**7.6**	**6.6**
Read instructions	8.8	8.8	8.8
Leave device to warm up	**N/A**	**4.6**	**4.6**
Read instructions	7.1	7.1	7.1
Take device from tray	6.8	6.8	6.8
Inspect expiry date	7.0	7.0	7.0
Check drug quality	**7.1**	**7.1**	**4.9**
Check medicine level	**7.1**	**N/A**	**N/A**
Select injection site	5.5	5.5	5.5
Bare injection site	5.0	5.0	5.0
Clean site with wipe	9.1	9.1	9.1
Remove and discard cap(s)	**6.3**	**6.1**	**6.1**
Pinch/stretch skin	5.5	5.5	5.5
Press device to skin	6.7	6.7	6.7
Press button/activate device	**9.9**	**9.9**	**7.3**
Wait until drug delivered. (Device A: check inspection window to see when injection is complete)	**8.5**	**5.2**	**5.2**
Remove from skin	4.7	4.7	4.7
Devices B and C: Check inspection window to confirm delivery of full dose	**N/A**	**4.7**	**4.7**
Dispose of device	5.0	5.0	5.0
Apply cotton wool to site	5.3	5.3	5.3
Dispose of other waste	6.5	6.5	6.5
Check for side effects (skin reaction)	4.6	4.6	4.6
Overall (just device use)	**11.5**	**11.5**	**10.4**
Overall (including device-independent steps e.g. cleaning the injection site)	**11.8**	**11.8**	**11.2**

4. Comparison of methods

4.1. Limitations of the user trials

4.1.1. Sampling issues

This study employed a quota sampling strategy in order to obtain a sample representative of the target population (see Section 2.3.1). As is common practice, age and gender were used to determine the quotas. Education level was also used in order to broaden the sample.

Some sample biases are observable, particularly towards more highly educated participants. These should be taken into account in interpreting the results. However, the sample did include at least eight people in each age category and five at each education level. This meets the recommendations of three to five users in each distinct group to discover the majority of usability problems (Nielsen, 2000; Caulton, 2001).

Moreover, the same trends in the measures were observable in all groups. For example, each of the groups found that device A took the longest to use, followed by B then C. This indicates that these trends are fairly reliable. However, the details may vary. For example, exact times may differ in a more balanced sample, so should not be used to set time limits for device use.

Age and education level are important but arguably a more important factor for inclusive design is the participants' capabilities. In common with many user trials, quotas were not set for capability directly. To enable investigation of this issue, participants were asked about their vision, dexterity and reach capabilities. These capabilities were chosen as the most relevant to autoinjector use. Cognitive capabilities are also important but less easy to measure in the time available. In addition, the devices are very unlikely to be prescribed to people with severe cognitive difficulties.

All participants had full vision ability (when using their normal glasses, if appropriate) and no difficulty reaching both arms out in front. Two of the 30 participants had slight difficulty with fine-

finger manipulation and a different two had slight difficulty with picking up and carrying objects. These numbers are comparable to the proportions of people with reduced capability in the target population. In the adult population in Great Britain, about 3.5% have reduced vision and 1.9% reduced reach, 4.8% have some difficulty with fine-finger manipulation, and 5.7% with picking up and carrying objects, using the same measures of capability as in our survey (Waller et al., 2013; Engineering Design Centre, 2012). The sample is statistically reasonably representative.

However, it does not include individuals with reduced vision and reach, combinations of capabilities and higher levels of capability loss. The problem is that these make up a relatively small proportion of the population, and thus do not feature in a sample of this size.

It is important to remember that some of the target users would have these capability losses that are not covered in the sample. In particular, some of the target users would have greater reductions in dexterity due to arthritis. In terms of interpreting the results from the user trials, it is expected that people with lower levels of capability would take longer and make more errors. Furthermore, certain types of errors, of particular relevance to inclusive design, may not be identified, as discussed in Section 4.3.

4.1.2. Other limitations

The study used training devices and prototypes instead of real autoinjectors for reasons of practicality, safety and ethics. However, the lack of a needle removed important feedback on device activation. Moreover, in practice, fear of needles may make participants more cautious in their use of the devices. Alternatively, the anxiety may lead to greater likelihood of errors. In addition, the devices activated more easily on the injection trainer than on real skin, particularly if the skin was not pinched or stretched properly. Thus errors related to pinching the skin were not accurately reflected in the user trials (see Section 3.1.3). These limitations reflect common practical and ethical limitations in user trials. Some could be addressed with more extensive trials at a later stage of design, but this is not always feasible.

The instruction sheets were based on the original instructions which were written with different amounts of detail and clarity. This may have influenced participants' responses, especially as their comments in Section 3.1.5 indicate that the instructions affect usability and response to a device. However, it is difficult to separate the instructions from the device itself. If a device is simpler to operate, then its operation can be explained more simply, resulting in simpler and shorter instructions. But this is not the only factor, as the same device can be explained with different degrees of clarity, and thus care needs to be taken in interpreting the results.

Practicalities of the user trials also meant that some steps in the task analysis could not be included. Although these were not core tasks, they may have provided context which would affect participants' behaviour. In addition, the user trials cannot identify potentially serious problems with these omitted tasks.

4.2. Limitations of the exclusion calculations

The exclusion calculation method is currently under development. In particular, it is currently limited by the underlying DFS dataset (see Section 2.4.1). The DFS measured disability prevalence for planning welfare support, and thus examined levels of capability causing significant difficulty with everyday tasks. This is not ideal for examining exclusion from product use. It does not identify difficulties for people with no or mild impairments, and is too coarsely grained to pick up fine differences in difficulty. As a result, the exclusion calculations may underestimate exclusion or fail to identify differences between devices.

Although the DFS is not ideal, it is the best dataset currently available to predict exclusion for tasks that involve several capabilities (Johnson et al., 2009). A survey is currently being planned to gather capability data specifically for exclusion calculations (Tenneti et al., submitted for publication).

4.3. Comparison of usability issues discovered

Although user trials and exclusion calculations use different measures, they are both commonly used to uncover usability and accessibility problems. We therefore compared the methods on the basis of the usability issues they discovered.

The user trials and exclusion calculations focused on different aspects of device use, and hence often identified different usability issues. In particular, the user trials identified unexpected or unusual task sequences that may not be adequately explored when working through a structured task analysis. On the other hand, the exclusion calculations identified problems for people with low capabilities who are often inadequately represented in user samples (Section 4.1.1).

For example, the user trials found that some people activated the device while off the skin. However, the exclusion calculations focused on the correct task sequence and so primarily examined pressing the device to the skin separately from pressing the button. Thus the possibility of activating the device off the skin was not thoroughly explored. The user trials also found that many users forgot to pinch or stretch the skin or did so wrongly. The exclusion calculations did not identify high demands here because they focused on the capability requirements of the action itself, not on whether people remembered what to do.

On the other hand, the exclusion calculations identified very high demands in cleaning the injection site, highlighting this as a key cause of exclusion overall. The user trials did not find this task problematic, probably because the sample did not include low enough dexterity and vision levels. Similarly, the exclusion calculations identified problems reading the print size in the instructions and handling the thin paper, in addition to difficulty understanding the instructions themselves.

This matches previous findings comparing heuristic evaluation (another usability inspection method) with user testing. Law and Hvannberg (2002) summarised previous work, noting that these methods often identify distinct sets of usability problems (with some possible convergence); and that user testing can provide deeper insight into problems.

Another finding was that the user trials were more effective at differentiating between devices on cognitive aspects. The exclusion calculations found fairly constant thinking demands across the devices (see Table 2), while the user trials found different levels of confusion and cognitive errors, particularly with the instruction sheets. This is because the cognitive scales currently used in the calculations are fairly coarsely grained and do not examine issues for people with no or mild impairment, as discussed in Section 4.2.

It is important not just to identify usability issues, but also to prioritise them so that redesign effort can be allocated appropriately. Exclusion calculations are particularly helpful here as they indicate how many people would be affected by an issue in the population as a whole. It is harder to extrapolate from user trials unless the sample size is very large.

4.4. Other contributions of the methods

Timings in user trials are often used to indicate usability, since easier devices are often faster to use. However, care is needed when using them in inclusive design, due to increased variation in the population. Furthermore, some older and less able users may prefer

to take longer and be less rushed, and this does not necessarily indicate usability problems. Timings are still useful, but need to be interpreted more cautiously and in conjunction with other measures.

Workload scores, such as TLX, indicate how hard participants felt they had to work when using a device (Hart and Staveland, 1988). They thus provide insight into the user's experience with the product. Furthermore, high workload often corresponds to low usability. Note that users with lower capabilities may have different expectations of the work required to use everyday products, and hence different perceptions of workload. Nevertheless, the scores are still useful for comparing devices and understanding users' perceptions of products.

The TLX scales include measures of perceived mental and physical demand. The TLX mental demand scores varied considerably across devices (Section 3.1.4), contrasting with the thinking demands from the exclusion calculations, which were identical for all devices. This agrees with the observation in Section 4.3 that the calculations are less effective at identifying differences in cognitive demands. There is a better match for physical demands with both methods indicating similar demands for devices A and B and lower demands for device C.

In summary, the exclusion calculations give a more detailed breakdown of demands, while the TLX scores have more discriminating power for mental demand. Further, they indicate perceived workload, rather than actual product demand. Both are important for understanding response to and use of products.

User preferences are also important for understanding user response, which is often based on other factors as well as usability. Furthermore, they enable users to highlight issues that are important to them, and can provide valuable suggestions for improving devices.

The population exclusion figures from the exclusion calculations are useful for assessing the numbers affected by design issues and determining if changes are worthwhile. They can also identify where improvements may fail to have the expected impact because other aspects of device use are still problematic. For example, in our study, changes to the device only reduced exclusion slightly because people were still excluded by the demands of cleaning the injection site (Section 3.2.4). This can provide valuable insight into design priorities that can be easily missed in user trials.

5. Recommendations

This study indicates that user trials and exclusion calculations complement each other, focussing on different aspects to give a fuller picture of the usability of a device in an inclusive design context. Furthermore, performing the trials and calculations as part of a single package improved the individual methods. In particular, the detailed task analysis needed for the exclusion calculations also underpinned the user trials, driving a greater degree of rigour. On the other hand, the exclusion calculations were informed by the preliminary results of the user trials, which highlighted some issues to look out for. This supports the recommendation by Clarkson et al. (2007) that these methods should be used together to provide the different kinds of guidance needed for inclusive design.

It is also important to consider how each method can be individually improved as it is not always possible to perform both. There may be insufficient time or budget for a full set of user trials. They are also difficult early in the design process, before a working prototype is available. Alternatively, sufficiently experienced personnel may not be available to perform reliable exclusion calculations.

One problem with user trials is that they may fail to identify potential issues for people with reduced capability. Many studies, like ours, seek to obtain a varied sample through quota sampling with variables such as age and gender. However, this is not enough to ensure that people with reduced capabilities are included.

Some authors recommend addressing this by involving users with more extreme needs or disabilities, and boundary users, who are on the boundary of being able to use the product (Clarkson et al., 2007). However, it can be difficult to identify who the boundary users are. One common strategy is to use older participants because they often struggle with product use. However, our study indicates that this is inadequate, as many older participants have high levels of capability. An initial analysis of appropriate boundary users for the product may help. For example, using the autoinjectors involved a lot of manual dexterity and vision, so an analysis would indicate that users with low dexterity and vision should be involved.

This may make recruitment more difficult. Furthermore, even with such efforts, it is often impossible to include enough people with each capability loss to ensure that all problems are identified. Focussing on extreme and boundary users can also skew the sample. This can be unhelpful for a mainstream product where designers are trying to be more inclusive, but not at the expense of the main user group. Therefore, it is still important to complement user trials with exclusion calculations, if possible.

User trials are also limited in the range of tasks that can be feasibly included, as some tasks are impractical to replicate. Exclusion calculations help designers to examine a wider range of tasks and to think more holistically about product use. If such calculations are not feasible, then designers should take other steps, such as a full task analysis, to ensure that wider product use is considered.

On the other hand, the study found that exclusion calculations may fail to identify some unexpected user actions. An improvement may be to broaden the initial task analysis beyond the "correct" sequence of product use, and consider a wider range of possible failures. Some observations of user behaviour would also help, even if full user trials are not feasible. Once identified, these unexpected behaviours can be explored more rigorously through analytic methods if necessary.

The study also highlighted the weakness of exclusion calculations in identifying cognitive issues. There are some cognitive appraisal methods (e.g. HEART: Williams, 1988) that fit well with the detailed task breakdown used in the exclusion calculations. These may work well together with the calculations to provide an insight into cognitive issues. However, they come from a different domain so further work is needed to investigate their effectiveness for inclusive design. Better underlying cognitive data would also help to improve this aspect of the calculations (see Section 4.2).

6. Conclusions

This study examined the effectiveness of two methods that are currently used in inclusive design: user trials and exclusion calculations. Both methods were used to examine three autoinjectors. The study found that each method, on its own, failed to identify all the usability problems. In contrast, when used together, the methods complemented each other. Each provided insight that the other lacked, as well as improving the execution of the other method. In particular, the user trials were more effective at identifying unexpected user actions and cognitive issues, while the exclusion calculations were better at identifying problems for people with low capabilities. The calculations also focused attention on the "highest hurdles": those aspects of product use that really limit the numbers who can use it, no matter how inclusive other aspects are. Together with the population exclusion figures,

this can help designers to prioritise design effort where it will really make a difference.

The paper also provides recommendations for how the individual methods can be used more effectively in inclusive design. However, the methods should ideally be used together to ensure a wide range of inclusive design issues are covered.

More case studies are needed to fully evaluate the contributions of different methods in inclusive design. Further work is also needed to improve and assess the methods based on the study findings. In particular, a survey is being developed to gather data more suited to exclusion calculations (Tenneti et al., submitted for publication). This should particularly improve the examination of cognitive issues, but work is needed to assess its effectiveness in practice.

Acknowledgements

The study was funded by Oval Medical Technologies. Further funding for considering the research implications was provided by EPSRC. We would also like to thank all who took part in and helped to run the usability trials.

References

Abbott Laboratories, 2013. Injecting with the HUMIRA® Pen. Website http://www.humira.com/global/injecting-humira-pen.aspx (accessed June 2013).

Amgen Inc., 2013. Injection Options and Demos. Website: http://www.enbrel.com/inject-ENBREL.jspx (accessed June 2013).

Byers, J.C., Bittner Jr., A.C., Hill, S.G., 1989. Traditional and raw task load index (TLX) correlations: are paired comparisons necessary? In: Mital, A. (Ed.), Advances in Industrial Ergonomics and Safety I. Taylor and Francis, pp. 481–485.

Cardoso, C., 2005. Design for Inclusivity: Assessing the Accessibility of Everyday Products. University of Cambridge, UK. Ph.D. thesis.

Cardoso, C., Keates, S., Clarkson, P.J., 2005. Are users necessary for inclusive design?. In: International Conference on Engineering Design (ICED) 2005, Melbourne, Australia.

Caulton, D.A., 2001. Relaxing the homogeneity assumption in usability testing. Behav. Inf. Technol. 20 (1), 1–7.

Clarkson, J., Cardoso, C., Hosking, I., 2007. Product evaluation: practical approaches. In: Coleman, R., Clarkson, J., Dong, H., Cassim, J. (Eds.), Design for Inclusivity. Gower, Aldershot, UK, pp. 181–196.

Combe, N., Harrison, D., Craig, S., Young, M.S., 2012. An investigation into usability and exclusivity issues of digital programmable thermostats. J. Eng. Des. 23 (5).

Ebling, M., John, B., 2000. On the contributions of different empirical data in usability testing. In: The Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques. ACM Press.

Engineering Design Centre, 2012. Exclusion Calculator. Available on the Inclusive Design Toolkit: http://www.inclusivedesigntoolkit.com/exclusioncalc/exclusioncalc.html (accessed June 2012).

Goodman-Deane, J., Langdon, P., Clarkson, P.J., 2010a. Key influences on the user-centred design process. J. Eng. Des. 21 (2–3).

Goodman-Deane, J., Ward, J., Hosking, I., Clarkson, P.J., 2010b. Comparison of the Usability of Three Autoinjectors. Technical report CUED/C-EDC/TR147. Department of Engineering, University of Cambridge.

Grudin, J., 2006. Why personas work: the psychological evidence. In: Pruitt, J., Adlin, T. (Eds.), The Persona Lifecycle, Keeping People in Mind Throughout Product Design. Elsevier, Amsterdam, pp. 642–663.

Grundy, E., Ahlburg, D., Ali, M., Breeze, E., Sloggett, A., 1999. Disability in Great Britain. Corporate Document Services, London. Research Report 94.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock, P., Meshkati, N. (Eds.), Human Mental Workload. North Holland B.V., Amsterdam, pp. 139–183.

Johnson, D., Clarkson, P.J., Huppert, F., 2009. Capability measurement for inclusive design. J. Eng. Des. 21 (2–3), 275–288.

Keates, S., Clarkson, J., 2003. Countering Design Exclusion: An Introduction to Inclusive Design. Springer, London.

Kirwan, B., Ainsworth, L.K., 1992. A Guide to Task Analysis. Taylor and Francis.

Klein, J.A., Karger, S.A., Sinclair, K.A., 2003. Digital Television For All: A Report on Usability and Accessible Design. Department of Trade and Industry, London.

Law, L.-C., Hvannberg, E.T., 2002. Complementarity and convergence of heuristic evaluation and usability test: a case study of UNIVERSAL brokerage platform. In: NordiCHI (The Second Nordic conference on Human–computer interaction). ACM Press, Aarhus, Denmark.

Nielsen, J., March 19, 2000. Why You Only Need To Test With Five Users. Jakob Nielsen's Alertbox. Available at: http://www.useit.com/alertbox/20000319.html (accessed June 2013).

Nielsen, J., Landauer, T.K., 1993. A mathematical model of the finding of usability problems. In: Conference on Human Factors in Computing Systems (CHI) '93. ACM Press, Amsterdam.

Oval Medical Technologies. Autoinjector technology. Website: http://www.ovalmedical.com/technology (accessed June 2013).

Persad, U., Langdon, P., Clarkson, P.J., 2007. Characterising user capabilities to support inclusive design evaluation. Universal Access Inf. Soc. 6 (2), 119–135.

Sims, R., 2003. 'Design For All': Methods and Data to Support Designers. Loughborough University, Loughborough. PhD thesis.

Tenneti, R., Langdon, P., Waller, S., Goodman-Deane, J., Ruggeri, K., Clarkson, P.J., Huppert, F.A. Design and delivery of a national pilot survey of capabilities. Int. J. Hum. Factors Ergon., submitted for publication.

Waller, S., Goodman-Deane, J., Langdon, P., Johnson, D., Clarkson, P.J., 2009. Developing a method for assessing product inclusivity. In: ICED '09 (International Conference on Engineering Design). Stanford, CA, USA.

Waller, S.D., Langdon, P.M., Clarkson, P.J., 2010. Using disability data to estimate design exclusion. Universal Access Inf. Soc. 9 (3), 195–207.

Waller, S.D., Bradley, M.D., Langdon, P.M., Clarkson, P.J., May 2013. Visualising the number of people who cannot perform tasks related to product interaction. Universal Access Inf. Soc..

Williams, J.C., June 1988. A data-based method for assessing and reducing human error to improve operational performance. In: IEEE Fourth Conference on Human Factors and Power Plants, pp. 436–450.