DEVELOPING A METHOD FOR ASSESSING PRODUCT INCLUSIVITY

Sam Waller[1], **Joy Goodman-Deane**[1], **Pat Langdon**[1], **Daniel Johnson**[2]
and P John Clarkson[1]
(1) Engineering Design Centre, University of Cambridge, UK
(2) Department of Psychiatry, University of Cambridge, UK

ABSTRACT

In order to develop more inclusive products and services, designers need a means of assessing the inclusivity of existing products and new concepts. Following previous research on the development of scales for inclusive design at University of Cambridge, Engineering Design Centre (EDC) [1], this paper presents the latest version of the exclusion audit method. For a specific product interaction, this estimates the proportion of the Great British population who would be excluded from using a product or service, due to the demands the product places on key user capabilities. A critical part of the method involves rating of the level of demand placed by a task on a range of key user capabilities, so the procedure to perform this assessment was operationalised and then its reliability was tested with 31 participants. There was no evidence that participants rated the same demands consistently. The qualitative results from the experiment suggest that the consistency of participants' demand level ratings could be significantly improved if the audit materials and their instructions better guided the participant through the judgement process.

Keywords: Inclusive design, capability data, product assessment

1   INTRODUCTION

Many people find products and services difficult to use. Some find them frustrating, some struggle but can manage if necessary, and others are effectively excluded from using them altogether [2]. There are many reasons for these difficulties, but a key root cause is that many products and services do not match the users' characteristics, capabilities and ways of thinking [1]. Inclusive design seeks to address this by making mainstream products usable by as many people as reasonably possible, without requiring them to use specialised adaptations [2].

However, designers often struggle to put inclusive design into practice. A recent survey found that a key factor is a lack of knowledge and tools that can motivate and equip them to carry out inclusive design effectively [3]. In particular, there is a need for methods that can evaluate the inclusivity of products and identify and prioritise how they could be improved. While several such methods do exist (see Section 2), few of these take a population perspective, which is required to address the need identified by Dong et al. for "methods of estimating the number of customers that might be included or excluded by the outcome of a particular design decision or by an existing design solution" [4].

The i~design research program [5] has therefore developed a method for evaluating the inclusivity of a product or service that is based on population data [1]. An exclusion audit estimates the proportion of a target population who would be unable to use a product due to the demands it places on key user capabilities. In a commercial context, and over a ten-year period, the EDC researchers have used a complementary trio of exclusion audits, user trials and expert appraisals to provide actionable insights for industrial partners, e.g. [6].

A critical component of an exclusion audit involves a demand assessment, where the assessor uses his or her judgement to compare the tasks involved in using the product with generic population data. For example, population data is available for the number of people who are unable to read a newspaper headline, large print text, and ordinary newsprint, and the assessor has to use this information to predict the number of people who would be unable to perform the vision related actions associated with using the product.

The degree of training required to perform exclusion audits limits their applicability, so as a first step towards enabling practicing industrial designers to perform these audits, the procedure for demand assessments was operationalised. An experiment was then designed and conducted to examine a specific aspect of the operationalised procedure, namely the consistency with which different participants judged the difficulty levels of an identical set of demands. This paper firstly presents the supplementary materials and the operating procedure that were developed to enable untrained participants to perform demand assessments. The specific actions that participants carried out are then described, followed by the results from the experiment and a discussion.

2   RELATED WORK

Many methods have been proposed for assessing the inclusivity of products and services [7-9]. Many of these involve users directly, obtaining their feedback on how they use the product in practice and the problems they experience. Such methods are extremely valuable in identifying problems that real users have in practice, yet the insights are inevitably drawn from a very small proportion of the target users, so it can be challenging to prioritise the significance of such insights in terms of the entire user population.

Analytical methods can help to address these shortcomings. They help a designer or usability expert to assess a product or concept, taking into account the issues that a range of users may experience. These methods also have limitations, primarily because they do not take advantage of the direct experiences of real users. However, they provide an important complement to the insights from user involvement: they can provide a population perspective, prioritising issues based on their significance according to the diversity that the experts are aware of within the population of interest [10].

Although there are many analytic methods available, such as task analysis and heuristic evaluation, most of these do not consider the whole range of diversity within the population nor do they directly relate their usability assessments to population figures. As a result, they do not consider important sections of the population, particularly for inclusive design. It can also be hard to prioritise these problems according to how many people would be affected. For example, designing to accommodate one area of impairment may impact on other aspects of the design, making it less usable by others [11]. An exclusion audit seeks to address this by systematically considering the whole range of capability losses evident across a national population. It can also provide figures indicating by how much alternative design improvements could reduce the exclusion.

Another design approach that uses population data on capabilities is engineering anthropometrics [12, 13]. However, this is rarely accompanied by a systematic method of assessing a product and it separates out the various measurements, examining just one capability (or part of a capability) at a time. Anthropometric methods are therefore often unsuitable to predict exclusion when several capabilities are used in combination to perform a task. For example pressing a button on a product may require hand-eye co-ordination, so the number of people who would be unable to press such a button cannot be ascertained from separate sources of vision and dexterity data, as the use of separate data sources will inevitably involve an unknown amount of double counting [1].

The HADRIAN tool [9] enables a designer to specify a task sequence and then examines which individuals in an anthropometric database would not be able to complete that sequence. However, the database is not intended to be representative of the target population and hence HADRIAN does not calculate national exclusion figures, which can help designers to better understand the true impact of design decisions.

An exclusion audit aims to provide a systematic way of assessing product use, matched together with a single source of generic capability data that can be applied to predict people's context dependent ability to perceive, think and act with real-world products. However, Johnson et al [14] found that no currently available UK dataset meets these criteria, and the best dataset is the Office of National Statistics 1996/97 Disability Follow-up Survey [15, 16]. This survey was commissioned to measure the prevalence of disability in Great Britain, in order to plan welfare support. Approximately 7200 participants were asked whether they could perform a series of everyday tasks, and their answers were used to estimate their level of ability in 13 different categories. Seven of these categories were selected for the exclusion audit as being most relevant for product use, as described in Section 3.1. However, the survey dataset is not ideal; for example, it was not designed for assessing products, has an untested theoretical basis, sampling biases and is based on self-report data. Nevertheless, this dataset remains the best available way to understand the prevalence for combinations of impairments, and to predict

how many people might be unable to perform tasks that involve several capabilities [14]. Furthermore, using this dataset allows initial versions of the exclusion audit method to be developed and tested, which can then be adapted to better data sources as they become available.

The exclusion audit method is based on a resource economic model as proposed by Kondrasake [17], which examines the match between the user's capabilities and the task's demands. Users are predicted to fail to achieve a task if any of their capabilities are less than that demanded by the task. Thus, if they have sufficient vision ability but insufficient hearing ability, they will still fail on the task. This was applied to product use by Persad et al [18], who laid out the theoretical basis for the exclusion audit method and outlined a new framework for future exclusion analyses.

An initial version of the exclusion audit method, using data from the Disability Follow-up Survey, was described in [1] and used successfully in both research and commercial contexts [6]. Building on this, Cardoso developed a more rigorous assessment process, incorporating a structured task analysis that encouraged consideration of the different capabilities that were employed [19].

More recently, the descriptions of the capabilities used in the method have been simplified. The original data used between five and thirteen scale points to describe a person's ability, with different numbers of scale points for different capabilities [15]. However, this was found to be difficult to use in the context of evaluating a product. The descriptions were therefore simplified to use a smaller and more consistent set of points [20]. Further work by Waller et al considered how to calculate the exclusion associated with tasks that may be achieved in several different ways [20], and how to visualise the results of an exclusion audit when multiple tasks make demands on multiple capabilities [21].

3   PROCEDURE FOR DEMAND ASSESSMENTS

Performing a demand assessment requires specifying the activities that are being assessed, then recording the particular demands that these activities place on the user's capabilities. The assessor then uses his or her judgement to compare each of these demands with the available population data, which is currently the Disability Follow-up Survey. In order to operationalise this process, the authors have developed supporting paper materials and a standard operating procedure, which are now described in turn. The paper materials include a booklet of standardised tasks for each capability and a worksheet that captures the task analysis and demand assessments.

3.1 Booklet of standardised tasks for each capability

The Disability Follow-up Survey covered many aspects of capability that are relevant for interacting with products, specifically: seeing (renamed to vision), hearing, dexterity, reach & stretch, locomotion, spoken communication and intellectual function (renamed to thinking). The process of converting the survey data into a format suitable for assessing products is now explained in further detail.

For each capability other than thinking, the Disability Follow-up Survey used a scale to measure the quality of life impairment associated with each survey participant's ability in that category [22]. A person with full ability has no quality of life impairment, while a person with very low ability will have severe quality of life impairment. Currently, the exclusion audit has to assume that the same scales can predict how well each of the survey participants would be able to use that capability to interact with the product being assessed. This assumption has some face validity, as a person with severe quality of life impairment will generally have a poor ability to interact with a product, however the specific mapping between these two types of scales is unknown, and in particular there is no guarantee that an incremental level on one scale will correspond with an incremental level on the other. Nevertheless, the assumption has to be made, as the Disability Follow-up Survey remains the best available dataset [14]. These quality of life scales will therefore now be referred to as ability scales.

Each ability scale was created from a different number of questions, although there were typically about 10 questions within each scale. The questions within each ability scale were examined to produce a consistent set of standardised tasks at four levels of demand, namely "No demand", "Low demand", "Moderate demand" and "High demand", as described in Waller at al. [20]. The complete set of standardised tasks is available online [2].

For each capability other than thinking, the standardised tasks form a scale that intends to measure how difficult a product is to use, according to the demands made on that particular capability. The standardised tasks were collated together to form a printed A5 booklet, with one page used to present

the standardised tasks for each separate capability. A further two-point version of the booklet was also developed, using only the "No demand" and "High demand" levels. This was compared to the 4-point version in the experiment.

The Disability Follow-up Survey data for thinking ability has a different underlying structure than all the other abilities, so assessing the demands that a product makes on the user's thinking ability requires a different procedure than the other demands. Subsequently, the thinking related data was actually printed on the participant's worksheet, rather than in the booklet, as this made it easier to keep track of the participants' responses.

3.2 Worksheet for task analysis and demand assessments

A small part the worksheet is shown in Fig. 1. The primary working space is a grid, where the top row specifies the tasks that are being assessed, and underneath this row, each cell contains the specific demands made by these tasks on each of the seven separate user capabilities. Also, for each capability other than thinking, a scale is printed within each cell, which can be used to rate the difficulty level of that demand, based on the standardised tasks from the Disability Follow-up Survey that are presented in the booklet. On this scale, N, L, M and H reflect "No demand", "Low demand" "Moderate demand" and "High demand" respectively. A further two-point version of the worksheet was also developed, using only the "No demand" and "High demand" levels, to correspond with the two-point version of the booklet.
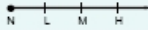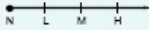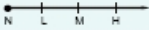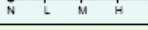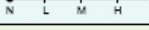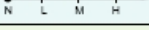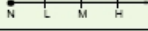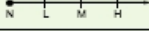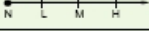

Figure 1. A portion of the grid used to record the tasks and demands associated with a product's use.

The Disability Follow-up Survey's thinking ability scale was constructed differently to all the others. The underlying questions used to construct the scale did not follow a graded order of difficulty, so each survey participant's thinking ability was calculated according to the number of everyday thinking tasks they could perform, from the list shown in Table 1. The same set of tasks were therefore printed in the bottom left corner of the worksheet, so that the assessor can mark each of the tasks that they consider relevant to achieving the goal with the product.

It should be remembered that the tasks chosen to measure the thinking ability, and the process of counting up the number of tasks to measure this ability were produced to assess the impact of disability on quality of life [22]. Using a similar process to measure the thinking demand imposed by a product may seem extraordinary, yet any predictive model for estimating design exclusion must consider the whole set of sensory, cognitive and motor abilities, which is currently best supported by the Disability Follow-up Survey capability database [16] in its original form. Developing and testing

the audit procedure with the existing data continues to help improve the procedure for such a time when better data is available.

Table 1. List of standardised thinking tasks

- Hold a conversation without losing track of what is being said
- Do something without forgetting what the task was whilst in the middle of it
- Think clearly, without muddling thoughts
- Count well enough to handle money
- Tell the time of day, without any confusion
- Watch a 30 minute TV program, and tell someone what it was about
- Read a short newspaper article
- Write a short letter to someone without help
- Remember a message and pass it on correctly
- Remember to turn things off, such as fires, cookers or taps
- Remember the names of friends and family that are seen regularly

3.3 Operationalised procedure for assessing demands

The operationalised procedure for performing demand assessments is described below. The output of this procedure is a quantitative description of the minimum level of vision, hearing, thinking, communication, locomotion, reach & stretch and dexterity abilities that are required to perform the task, were each level is directly compatible with the measures used within the Disability Follow-up Survey.

3.3.1 Specify assumptions

The assessor first chooses the product and goal being assessed, and records these on the worksheet. The goal here refers to the task that is to be achieved with the product, such as boiling water using a kettle. The initial state of the product and relevant items, and aspects of the environment will affect a person's ability to use the product successfully. The assessor therefore also records any assumptions made about the initial state and the environment.

3.3.2 Perform a task analysis

The assessor then identifies one common way of achieving the goal, and uses the top row of the worksheet to describe this sequence of tasks, as shown in Figure 1. These tasks should be chosen to communicate how the goal is achieved in an unambiguous yet concise manner. The version of the method tested in this study examines a single way of achieving the goal, using a single linear sequence of tasks, with a single assumed environment and initial state. Waller et al [20] discuss some of the issues involved in extending this to deal with alternative task sequences.

3.3.3 Identify the demands associated with the tasks

The assessor then examines each task in turn. For each task, he or she identifies and records the actions that place demands on each of the seven different user capabilities measured within the Disability Follow-up Survey. A demand is usually expressed using a verb that relates to the capability (e.g. see, read, perceive) and a noun, specifying the part of the product (or other item) that is used (e.g. handset, button, dial tone). Each demand should usually refer to a single product feature, and some examples are shown in Fig. 1. If a task places no demand on a capability, then the assessor draws a line through the corresponding box.

Reach & stretch and thinking demands are treated slightly differently. Reach & stretch demands describe the position of the users' hands in relation to their bodies, which may be simply conveyed using a sketch, also shown in Fig. 1. The thinking demands associated with achieving the goal may include actions associated with the specific tasks, such as recognising symbols, in addition to actions involved with the process as a whole, such as maintaining concentration and planning ahead. Both the specific and overall types of demand should be recorded on the worksheet.

3.3.4 Rate the difficulty level of each demand

For each capability other than thinking, the assessor then rates the difficulty level of each of these demands, by comparing them with the set of standardised tasks for that capability, which are printed in

the supplementary booklet. Note that difficulty is used here to mean the general difficulty of the task, based on how that capability varies within the UK population.

This difficulty should be assessed by judging whether the task at hand is harder, easier or about the same as the standardised tasks (by trying out both the product and the standardised tasks, if possible). The assessor marks the corresponding scale within the worksheet to reflect the assessment. Figure 2 shows an example difficulty assessment, where seeing the handset was judged to be slightly easier than recognising a friend at arm's length away, which is one of the statements at "Low demand".
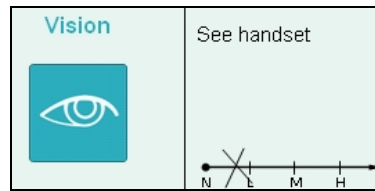

Figure 2. Example of a difficulty level assessment.

The assessor then rates the overall difficulty level. This should usually be the maximum difficulty level in that capability, because a user who is excluded from any part of the product use is excluded from the whole. However, a higher rating may be given, if the assessor feels that fatigue or other effects make the combined demand greater than any individual demand.

The Disability Follow-up Survey's thinking ability scale was constructed differently to all the others, so in an exclusion audit, the assessor considers both the thinking demands associated with specific tasks, and the overall thinking demands associated with achieving the goal, and then uses the worksheet to mark each of the tasks in Table 1 that are considered relevant to achieving the goal. The thinking demand is then summarised by counting up the number of tasks that were marked.

Now that all the demands have been compared against the standardised tasks, the overall difficulty levels on each capability could be used within a simple algorithm to predict how many of the 7200 Disability Follow-up Survey participants could not perform the task [20]. Given that this survey is nationally representative, it is then possible to predict the proportion of the UK population that would be excluded from the task. However this paper focuses on the consistency with which different assessors judged the difficulty levels of the same set of demands, so the use of these difficulty levels to calculate exclusion is not considered in any further detail.

4 EXPERIMENT DESIGN

4.1 Aims

The experiment focused on one key part of the exclusion audit method: rating the difficulty levels of the demands (as described in Section 3.3.4). Estimating the proportion of a national population that will be unable to use a product will always be likely to involve a judgement process to compare the tasks at hand against generic capability data. The consistency of these judgements across different assessors is critical (although not sufficient) to ensure validity of the method as a whole. Testing and improving the consistency of these judgements is therefore an important first step in developing the method.

The experiment was designed to investigate the consistency in participants' difficulty ratings. Two products were assessed, to ensure that all seven capabilities were involved and to aid the generality of the results. The authors therefore performed the first three stages of the operationalised demand assessment procedure, described in Sections 3.3.1 to 3.3.3, and developed corresponding partially completed worksheets, an example of which is shown in Fig. 1. The participants used these pre-completed worksheets to assess the difficulty levels of the demands, as described in Section 3.3.4. The experiment was also designed to examine whether the consistency of participants' judgement of the difficulty levels was affected by the use of two or four points to define the demand scales. Half of the participants therefore assessed the products using two-point scales throughout, while half the participants used four-point scales throughout, as described in Section 3.1.

4.2 Experiment procedure

The participants' consent was obtained in accordance with ethical guidelines. They were then given a brief introduction to the exclusion audit method, and were shown an example partially completed

worksheet for assessing a stapler, where the tasks and demands were specified in advance, but the scales for assessing the difficulty levels of these demands were unmarked (similar to the example in Fig. 1). The experimenter described the various parts of the form verbally, using a standard script, and explained that the participants would then assess the difficulty level of similar demands for two different products.

The participants were then given written instructions that explained in detail how to make the assessments, the accompanying booklet, a product to assess, and the corresponding pre-completed worksheet for that product. The instructions particularly emphasised the importance of trying out the tasks with the products, and making the difficulty assessments by comparison with the standardised tasks that were written in the booklet. Once the participant had finished an assessment with one product, they were offered an opportunity to take a break, before completing a similar assessment with a second product.

The participants assessed recharging batteries using a battery charger and phoning the speaking clock using an office telephone. The order of the products was counter-balanced, with half of the participants assessing the charger first, and half assessing the phone first. After assessing both products, participants completed a questionnaire, providing some demographic information and describing their thoughts about the method. Some of the participants were further interviewed about how they made their difficulty judgements.

4.3 Sample

Participants were recruited through university societies and a graduate news bulletin in the University of Cambridge, UK. Thirty-one people took part in the experiment (15 men and 16 women), most of whom were students (both undergraduate and postgraduate). This group was considered to have similar levels of education to the designers and design managers that the method is aimed at. The results from this first experiment enable the method to be improved, before it is tested a sample that better represents the intended target users, namely industrial designers. Each participant did not have any previous experience related to exclusion audits.

Three of the participants were discarded from the quantitative analysis, for the following reasons: one because the participant was ill during the experiment; one because the interview revealed that the participant had rated easy demands as "high" and hard demands as "low"; and one because the participant had assessed every task as "no demand", thus skewing the results. A slightly different set of three responses were discarded for analysis of the thinking demands: one because the participant was ill (as before), and two because the participants had completely misunderstood the task. The responses from the entire sample were included in the qualitative results, as they provide valuable insight into the use of the method. The breakdown of the sample is shown in Table 2.

Table 2. Breakdown of the sample by condition and analysis type

	2 point scales		4 point scales	
	Charger first	Phone first	Charger first	Phone first
Sample for quantitative analysis	7	7	7	7
Sample for qualitative analysis	8	8	8	7
	Charger first		Phone first	
Sample for analysis of thinking	14		14	

5   QUALITATIVE RESULTS

5.1 Likes and dislikes

When asked what they liked about the method, 15 participants referred to the instructions and/or the method being clear and easy to follow. Five people also thought the method was well structured and organised. Many participants also liked the colour coding and layout of the worksheet. Six commented that the method makes you consider a range of capabilities, with two people saying it made them think about aspects they had previously taken for granted, such as the need to use a variety of senses when using a product.

When asked what they disliked about the exclusion audit method or thought could be improved, the most common complaints were about the thinking assessments, mentioned by 15 participants. In particular, eight participants felt that the thinking instructions were confusing, and eight commented that the set of standardised tasks were hard to relate to the actual task. Other thinking related dislikes were identified by four people or less.

Three participants also said they had difficulty determining the order in which to do the assessments, three participants had difficulties performing ratings on a sliding scale, and four participants indicated varied dislikes about the design of the assessment sheet.

5.2 Problems with the procedure

The audit procedure depends upon determining how the product stands in relation to the standardised tasks in order to calculate exclusion figures. It is therefore vitally important that the difficulty assessments are made by comparing the product demands against the standardised tasks. This was explained several times in the instructions. However, many of the participants did not appear to follow this procedure.

During each trial, the experimenter took note of the extent to which each participant referred to the instruction booklet. Based on this data, about a third of participants (10) only looked at the booklet once, and did not refer to it again afterwards, when actually doing most of the assessments. Of these 10 participants, six were using the 2-point scales and four were using the 4-point scales. The authors consider it unlikely that these participants performed their assessment judgements against the standardised tasks, as this would have involved them remembering the entire booklet in sufficient detail to perform the comparison.

To investigate this further, the latter 17 participants were asked how they actually made their assessments. Of these, 10 said that they assessed the demands against the statements in the booklet, although in some cases this was from memory. The others said they based their assessments on their own level of difficulty or effort, or that of an average person, or more vague ideas of how hard the tasks were to do.

6 QUANTITATIVE RESULTS

6.1 Initial analysis of the difficulty assessments for all capabilities other than thinking

Participants rated the difficulty levels of each demand on a continuous scale: they could mark the scales anywhere between or above the fixed points (see Figure 2). To reflect this, their responses were measured in an integer number of millimetres along the scale, which was 21 mm long.

For each separate demand assessment, histograms were plotted for the 14 participants that used a 4 point scale, examples of which are shown in Figs 3(a) and (c). The histograms were also plotted for the 14 participants that used a 2 point scale, as shown in Figs 3(b) and (d). The participants overall demand assessments were not considered in any further detail, as these compound several sources of variability. The authors' judgement of the absolute minimum acceptable level of consistency for demand assessments is for the participants to agree to within any given band that covers one third of the scale, meaning that all assessments should be less than 7 mm apart.

To examine the consistency of participants actual demand assessment judgements compared to chance, imagine a scenario where 14 participants generated a random integer between 0 and 21 to determine the position in millimetres where they should mark the scale. A cumulative binomial calculation provides that the probability of 10 or more participants marking between 0 and 6 mm is 0.273%. Ignoring some instances of double counting, there are 16 other equivalent bands that participants could mark in to satisfy the condition of being less than 7 mm apart, so the probability of 10 or more participants' assessments being less than 7 mm apart is therefore less than 16*0.273 = 4.4%. The participants assessments of the actual demands were therefore placed into one of two categories, depending on the consistency of participants judgements:

- Some consistency: If 10 or more participants (71%) answered within 7 mm of each other, such as for Figs 3(a) and (b)
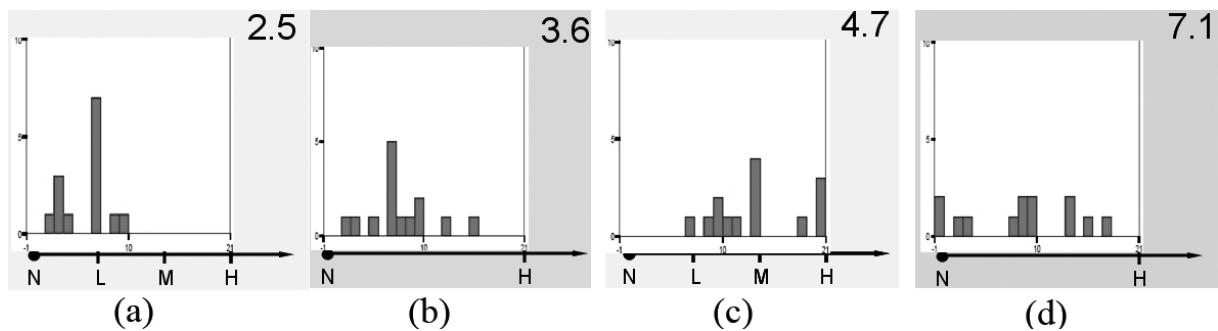- No consistency: Any remaining demands, such as Figs 3(c) and (d).

Figure 3. Example histograms of participants' demand assessments.
The printed numbers are the corresponding standard deviations in mm.

Considering both the charger and the phone, each participant assessed 42 separate demands. Of these assessments, the percentages that met the criteria for each agreement category are shown in Table 3, for the 2-point and 4-point scales. Although over half the demands did not receive consistent judgements for either type of scale, the judgements for the 4-point scales appear to be more consistent than the 2-point scales.

Table 3. Percentage of demands that were consistent, for each type of scale.

	2-point scale	4-point scale
Some consistency	12%	46%
No consistency	88%	54%

6.2 Consistency in prioritising tasks for capabilities other than thinking

A key purpose of an exclusion audit is to identify opportunities for improving the product or service. This primarily depends on the participants' assessments of the tasks or the product features that make the greatest demands on the user's capabilities, so the results were also analysed to examine the consistency of the participant's relative judgements of the highest demand made on each capability. The results are only presented for the 4-point scales, because Section 6.1 indicates that these scales had higher consistency. For each demand on each capability, the graphs in Figure 4 show the percentage of participants who thought that demand was the hardest, or equal hardest.

Many participants rated two or more demands as equally highest, so the sum of all the bars on each graph may exceed 100%. The graphs are only plotted for capabilities where participants were asked to rate more than one demand, because the participants will obviously always agree on the highest demand if there is only one demand.

Considering the vision demands for the phone, four separate demands were assessed, no participants thought that the first task was the hardest, and just over 80% thought the second task was the hardest (or equal hardest). Considering all the demands on both products, between 60% and 80% of participants agreed on which demand was the highest within each capability. However, this result has not been tested for significance and in some cases there were only four different tasks to choose from.

6.3 Analysis of thinking assessments

The thinking demands were assessed differently, as explained in Section 3.3.4. Instead of assessing the individual thinking demands, participants marked each of the 11 tasks shown in Table 1 that they thought were relevant to achieving the goal. To examine the consistency of the participants' judgements, we imagine a scenario where the 28 participants flipped a coin to choose if that task was relevant. This is equivalent to assuming that each judgement was made randomly and independently. In this scenario, there is a 96.4 chance that between 9 and 18 participants would mark the task. It is therefore considered that participants consistently judged a standardised thinking task as being relevant to achieving the goal if more that 18 participants marked it, and not relevant to achieving the goal if less than 9 marked it. The judgement is considered inconsistent if between 9 and 18 participants marked it.

For the charger, eight of the tasks were considered not relevant, and three tasks received inconsistent judgements. For the phone, one task was considered relevant, seven not relevant, and three tasks received inconsistent judgements. The task of "Do something without forgetting what the task was whilst in the middle of it" was the most frequently marked for both products. 22 participants marked

this task for the phone, which meets the criteria for a consistent judgement. 17 participants marked this task for the charger, which is classed as an inconsistent judgement.
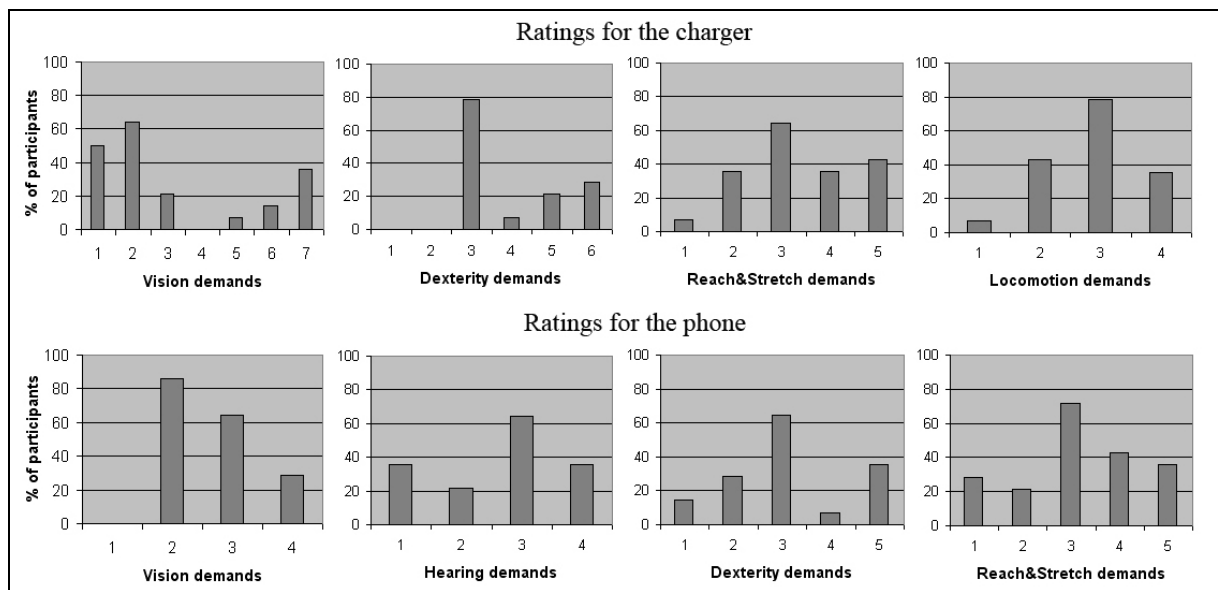

Figure 4. The demands rated highest by the 14 participants.

The task of "think clearly, without muddling thoughts" is one example of a task that received inconsistent judgements for both products, with 16 participants marking this task for the charger, and 12 marking it for the phone. The task of "Write a short letter to someone without help" was consistently judged as not relevant, with no participants marking this task for either product. Participant therefore seemed to agree that some of the tasks were definitely not relevant, but could only agree on one task being relevant for one product.

7   DISCUSSION

The exclusion audit is primarily intended to assist design commissioners and practicing design teams to evaluate prototypes. The authors have therefore planned a program of research that will culminate in assessing the validity of an exclusion audit with a sample that represents its intended users. However, as the authors had never previously written instructions for an exclusion audit that were intended for an untrained user, or measured the consistency of different assessors judgements of demand levels, it was considered prudent to first explore the utility of the exclusion audit with a more readily available sample, namely students.

The majority of participants' difficulty level judgements did not reach the level of consistency that the authors consider satisfactory, namely that participants should agree to within any given band that covers one third of the scale. Based on observations made during the experiment, the authors consider that the use of "Low", "Moderate" and "High" to describe the difficulty levels allowed participants the freedom to use their own internal definitions of these difficulty levels, in conjunction with or instead of the definitions given in the booklet.

For example, "High demand" could be taken to mean: the highest demand that the participant can imagine, the highest demand the participant thinks this product makes, or the participant's own perception of whether the task was difficult. Indeed, the participant might find any of these three options more salient than comparing against a particular definition of "High demand" printed within a booklet. Further evidence of participants misunderstanding the intention of "High demand" is evident from interviews that were conducted after the experiment, where several participants revealed that they judged reading the "+" sign on the charger as being harder than reading ordinary newsprint, which is the standardised task for "High demand". However, none of the participants marked the scale in a position above "High demand" to reflect this judgement. This issue may perhaps be resolved by increasing the training given to participants, using numbers to label the scales instead of words, or using software or flowcharts to help guide the participants' judgements.

The standardised tasks used within the Disability Follow-up Survey were never intended for the purpose of assessing products, so they may only reflect a subset of the factors that are needed to

predict someone's ability to interact with a product. This is particularly apparent for the thinking assessments, as it is not possible to construct a scale out of the corresponding standardised tasks, and many of these tasks seem unusual in the context of assessing products.

Within the i~design3 research program [5], research is currently underway to investigate what measures of capability, within the constraints of a national survey, would be most relevant to predicting a user's ability to perceive, think and act when interacting with products, and how an initially untrained assessor could best use these measures to provide a consistent and valid prediction of the number of people who would not be able to perform specific actions with a product.

8   CONCLUSIONS AND FURTHER WORK

In a commercial context, inclusive design researchers have successfully used a trio of exclusion audits, expert appraisals, and user trials to provide actionable insights for industrial organisations. As an initial step towards enabling practicing industrial designers to uncover these insights for themselves, the procedure to perform demand assessments within an exclusion audit has been standardised and developed, together with supplementary paper materials. A preliminary experiment examined one specific aspect of this procedure, namely the consistency with which different participants judged the difficulty levels of an identical set of demands.

Results indicate that participants liked the method and found that it helped them to consider a range of capabilities, as well as things they had previously taken for granted, such as the need to employ a variety of senses in using a product. However, their judgements were found to be inconsistent. Based on the qualitative results from this experiment, the authors conclude that the consistency of untrained users' judgements could be significantly improved if the audit materials and their instructions involved an unavoidable comparison against the standardised tasks, rather than using terms such as "low", "moderate" and "high" to represent the standardised tasks.

Nevertheless, although not statistically tested, participants' ratings appeared to be more consistent when using a 4-point scale than a 2-point scale, and 60-80% of participants tended to agree on the task or product feature that made the greatest demand within each capability.

REFERENCES
[1]   Keates S. and Clarkson J. *Countering design exclusion: An introduction to inclusive design*, 2003 (Springer, London).
[2]   Clarkson P.J., Coleman R., Hosking I. and Waller S. *Inclusive design toolkit*, 2007 (Engineering Design Centre, University of Cambridge). Avail. at www.inclusivedesigntoolkit.com [Accessed: May 09].
[3]   Goodman J., Dong H., Langdon P. and Clarkson P.J. Increasing the Uptake of Inclusive Design in Industry. *Gerontechnology*, 2006, 5(3), 140-149.
[4]   Dong H., Keates S. and Clarkson, P.J. UK and US industrial perspectives on inclusive design. In *Proc. Include 2003*, Royal College of Art, London, 2003.
[5]   i~design project website. www-edc.eng.cam.ac.uk/idesign3/ (accessed May 09)
[6]   Klein J.A., Karger S.A. and Sinclair K.A. *Digital Television for All: A report on usability and accessible design*, 2003 (Department of Trade and Industry, London). Avail at www.digitaltelevision.gov.uk [Accessed May 09]
[7]   Benyon D., Crerar A. and Wilkinson, S. Individual Differences and Inclusive Design, pp. 21-46. in Stephanidis C. (ed) *User Interfaces For All*, 2001 (Lawrence Erlbaum Associates).
[8]   Pirkl J. J. *Transgenerational Design - Products for an Aging Population*, 1994 (Van Nostrand Reinhold, NY).
[9]   Porter J.M., Case K., Marshall R., Gyi D. and Sims, R. 'Beyond Jack and Jill': designing for individuals using HADRIAN. *Int. J. Industrial Ergonomics*, 2004, 33(3), 249-264.
[10] Persad U., Langdon P. and Clarkson P.J. A Framework for Analytical Inclusive Design Evaluation. In *International Conference on Engineering Design, ICED 2007*, Paris, 2007.
[11] Newell A. F. and Gregor P. User Sensitive Inclusive Design. In *Proceedings of the 2000 conference on Universal Usability* pp. 39 – 44.
[12] Smith S., Norris B. and Peebles L. *Older Adultdata: The Handbook of Measurements and Capabilities of the Older Adult*, 2000 (Department of Trade and Industry, London).
[13] Stanton N., Hedge A., Brookhuis K., Salas E. and Hendrick, H. *Handbook of Human Factors and Ergonomics Methods*, 2004 (CRC Press).

[14] Johnson D., Clarkson P.J, & Huppert F. Capability Measurement for Inclusive Design. *Journal of Engineering Design*, to appear

[15] Grundy E., Ahlburg D., Ali M., Breeze E. and Sloggett A. *Research report 94: Disability in Great Britain*, 1999 (Corporate Document Services, London)

[16] Department of Social Security Social Research Branch, Disability Follow-up to the 1996/97 Family Resources Survey [computer file]. Colchester, Essex: UK Data Archive [distributor], 3 March 2000. SN: 4090

[17] Kondraske G.V. Measurement Tools and Processes in Rehabilitation Engineering. In Bronzino J.D. (ed) *The Biomedical Engineering Handbook*, 2000 (CRC Press, Boca Raton), Chapter 145.

[18] Persad U., Langdon P., and Clarkson P.J. Characterising user capabilities to support inclusive design evaluation. *Universal Access in the Information Society*, 2007, 6(2).

[19] Cardoso C. *Design for Inclusivity: assessing the accessibility of everyday products*. PhD thesis, 2005. Department of Engineering, University of Cambridge.

[20] Waller S.D., Langdon P.M. and Clarkson P.J. Using disability data to estimate design exclusion. *Universal Access in the Information Society*, 2009, to appear.

[21] Waller S., Clarkson J. and Langdon P. Visualising design exclusion predicted by disability data. *UAHCI 2009*, part of *HCI Internationl 2009*, San Diego, 2009, to appear.

[22] Martin J. and Elliot D. Creating an overall measure of severity of disability for the office of population census and surveys disability survey. *J. Royal Statistical Society Series A*, 1992, 155(1): 121-140

Contact: Dr Sam Waller
Engineering Design Centre
Department of Engineering
University of Cambridge
Trumpington Street
CB2 1PZ, UK
Tel: +44(0)1223766961
Email: sdw32@cam.ac.uk
URL: http://www-edc.eng.cam.ac.uk/people/sdw32.html

Dr Sam Waller is a researcher within the inclusive design group at the University of Cambridge, Engineering Design Centre. Sam's research investigates methods with which national population statistics for capability can be collected, and then how this information can be presented to calculate the number of people who would be excluded from using a certain product.