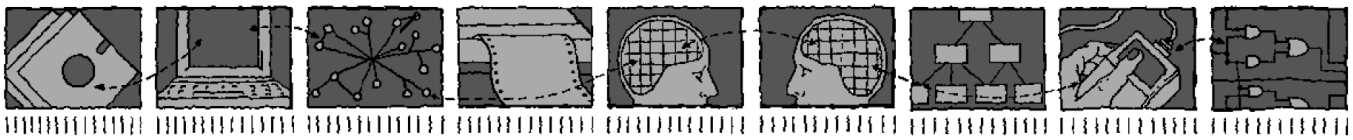
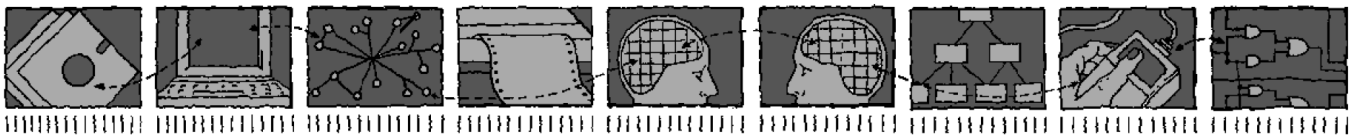


Department of Computing Science and Mathematics
University of Stirling



Can People Program Their Home?

Claire Maternaghan

Technical Report CSM-191

ISSN 1460-9673

April 2012

Department of Computing Science and Mathematics
University of Stirling

Can People Program Their Home?

Claire Maternaghan

Department of Computing Science and Mathematics
University of Stirling
Stirling FK9 4LA, Scotland

Telephone +44 1786 467 421, Facsimile +44 1786 464 551
Email cma@cs..

Technical Report CSM-191

ISSN 1460-9673

April 2012

Abstract

The requirements of a home system vary greatly in terms of the residents, the existing devices available, and many other factors. It is therefore crucial that any home system can be customisable by the end user; ideally this process would be simple and quick to perform. However, the challenges of end user programming are enormous. There is considerable challenge in obtaining concrete and unambiguous rules from non-technically minded individuals who typically think, and express, their desires as high level goals. Additionally, the language that is used to express the home rules must enable simple rules to be written by those less technical or ambitious, yet also not restrict more competent users.

Homer is a home system that has been fully developed by the author [4]. It offers the ability for users to control, monitor and program their home. A custom policy language has been designed called Homeric which allows rules to be written for the home.

In order to evaluate Homeric, and the design guidelines for allowing end users to formulate Homeric, a stand-alone application was written. This application is called the Homeric Wizard and allows rules to be expressed for the home. It makes use of natural language and visual programming techniques, as well as the notion of perspectives which allows rule elements to be browsed from differing aspects – devices, locations, people and time.

An online evaluation was performed over two weeks, obtaining participation from 71 individuals of varying age, gender, status and technical abilities. This report presents an evaluation of the Homeric Wizard tool, and all the results and hypotheses obtained.

Acknowledgments

Claire Maternaghan is supported by the Scottish Informatics and Computer Science Alliance, the University of Stirling, and the MATCH project (Scottish Funding Council, grant HR04016). The author is grateful to Kenneth Turner, Marilyn Lennon and Kate Howie for their advice and support, as well as aiding the distribution of the evaluation. Thanks also to all those who participated in the evaluation.

Contents

1 Introduction

A common limitation of current home automation systems is that the home logic is hidden, so that it cannot be controlled or changed by the user. This can result in the home behaving in ways which the user does not understand and cannot discover without contacting the system installer. It is also common to find that changes to the home logic have to be made by the system installer (at cost). Some commercial systems do allow the user to create rules for the home. However the user interfaces for this are often complex and hard to use, meaning that the average householder is unlikely to attempt any changes.

Homer is a home system which allows end users to fully program their home using a language called Homeric [4]. To evaluate some aspects of Homer an evaluation was carried out. This evaluation was designed to investigate two aspects of Homer. The first is the usablility of Homeric: can Homeric be understood by both technical and non-technical individuals, as well as used to express desired automation for the home? Secondly, some design guidelines for end user applications for the home were evaluated. In particular, the evaluation explored the notion of perspectives and the idea of using a hybrid of natural language and visual programming techniques to allow both technical and non-technical individuals to program their home. This evaluation was required to gauge the success of these aspects of Homer.

Section 2 describes the Homeric Wizard tool which was used to aid the evalutions. Section 3 describes the evaluation format and questions. The participants who took part are described in section 4. Section 5 reviews the results of the evaluation, looking at both the quantitative and qualitative data, and using the quantitative data to evaluate various hypotheses. Finally, the report concludes with a summary of the findings in section 6.

2 The Homeric Wizard

The Homeric Wizard tool is available to view and test at [6].

2.1 Goal

The Homeric Wizard was built solely as a tool to evaluate the Homeric language, perspectives and end user programming techniques. The wizard is a standalone prototype web-based application which can be customised to the individual.

2.2 Background

Having thoroughly researched the area of end user programming of the home [5], design guidelines and ideas have been produced [4]. These involve primarily two main concepts: using a hybrid of natural language and visual end user programming techniques, and the notion of perspectives. Each of these is now discussed in turn.

2.2.1 End User Programming Techniques

There exist four main end user programming techniques:

- Programming by demonstration: the system can monitor your behaviour to effectively program itself.
- Tangible Programming: the end user can program by manipulating physical objects.
- Natural Language Programming: through natural language an end user can express what they want, for example through speech or text, and the system can analyse this to extract the core program logic.
- Visual Programming: offers a pictorial means to program, such as through diagrams or icons.

The most successful existing tools which offer end user programming within the home generally make use of a hybrid of these techniques. Two examples include Alfred [3] (demonstration and natural language) and CAMP [8] (natural language and tangible). These projects offer a flexible solution, allowing the user to find ways that work best for them, rather than being forced into one particular programming method.

The Homeric Wizard tool makes use of both natural language and visual programming techniques in the hope of easing and simplify programming the home for the user. Natural language was chosen as it was felt to be the simplest of all techniques for the user, and relies far less on the user remembering particular symbols

or abreviations. However, words alone were considered too open-ended for both the user and the sytem, and visually less attractive and uninspiring. Hence, combining the familiarity and flexibility of natural language with the attractive but restrictive nature of visual programming results in a hybrid of simplicity and order.

2.2.2 Perspectives

The flaw with many existing end user programming solutions is the requirement for the user to conform to a particular way of thinking. This can be seen in systems such as iCap [7], Indigo (`www.perceptiveautomation.com`), Media Cubes [2] and SiteView [1]. Fellow researchers also agree, and user studies performed by Truong *et al.* [8] confirm these beliefs.

Homer makes use of the notion of persepectives, whereby users are able to locate various aspects and features within their home via four different categories (perspectives): location, device, time, and person. This allows for the varying ways users can visualise or refer to a particular device, trigger, condition or action to be supported, instead of forcing one particular way upon a user.

2.3 Design

The Homeric Wizard is designed to be as simplistic and visually appealing as possible. There exists only one main screen at all times, which allows one rule for the home to be expressed. This main screen, as seen in Figure 1, comprises a when panel, a do panel, and a library of triggers, conditions and actions which can be dragged and dropped over the desired when or do panel. Colour is used to indicate which term can be dropped on which panel.

Figure 1: The Homeric Wizard Main Screen - with overlay text

2.3.1 Rules

A rule is expressed as a series of triggers and conditions within the when panel, and a series of actions (with optional conditions) in the do panel. Triggers and conditions can be optional, required or ordered. The full Homeric language specification can be found in [4].

2.3.2 Customisation

There are three aspects of the wizard that can be customised: names, perspectives and difficulty levels. Each is discussed in turn below.

Names In order to provide a personalised experience, the user's name, partner's name and other members of the household can be optionally entered. These names will then be used in the naming of various devices and locations within the library of events. For example: "**Claire's** Bedroom", "**Claire's** Birthday", "Send email to **Claire**".

Perspectives It is possible to toggle perspectives on or off. When off, the list of device types is shown in the library. When on, four higher level menu categories are shown instead: Devices, Locations, People and Time. This allows for the same range of triggers, conditions and actions to be located through multiple menus. An example of this can be seen in Figure 2.

Figure 2: The Homeric Wizard Perspectives - with highlighted examples

Difficulty Levels There are three different difficulty levels supported by the wizard:

- Level 1 - Easy ("Keep it simple!"): Does not allow the user to group triggers and conditions in the when panel. Instead, all triggers and conditions are listed as required, with "and" joining them. They are unable to have conditions within the do panel.
- Level 2 - Intermediate ("I'd like to play!"): Allows groupings of triggers and conditions within the when panel, using "all of the following", "any of the following" and "all of the following in order" to indicate "and", "or" and "then" respectively. They are unable to have conditions within the do panel.
- Level 3 - Advanced ("Let me see it all!"): Allows users to have groupings of triggers and conditions within the when panel, as well as conditions within the do panel.

The difference between the three types of rules allowed with these difficulty levels can be seen in Figures 3, 4 and 5, which present a rule for automating part of a morning routine in each difficulty level.

2.3.3 Examples

To illustrate the Homeric Wizard tool, three example policies are presented in Figures 3, 4 and 5. Each example presents a rule which will perform some automation for a morning routine. The following examples also illustrate the customisation of the tool using names, as "Claire" appears throughout the examples.

The first of these rules is very basic, and can be written using any difficuly level. The left hand side presents the when panel, which specifies all the triggers and conditions that must take place for the rule to fire: in this case, when the alarm clock turns on in the morning. The right hand side presents the do panel, which specifies all the actions that must be performed when the rule is fired. In this case, the kettle should be turned on and the curtains in the bedroom should be opened.

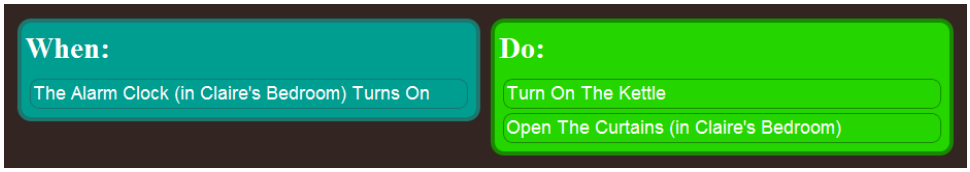


Figure 3: Sample Homeric Policy: Level 1 (easy)

The second of these rules involves groups of triggers and conditions in the when panel. This, therefore, means that users who are using difficulty level 1 would not be able to produce this rule. The when panel describes three triggers and conditions, and involves one sub-group. This therefore reads: "when it is a weekday and (the alarm clock turns on then the bed detects Claire gets out of bed)". As can be seen in the screenshot, grouping is represented visually by allowing terms to be dropped within the group. This group can then be dragged within the when clause to reorder it. Similarly, terms and groups can be removed by dragging the term/parent term (group) to the bin at the top of the screen (omited in this screenshot for simplicity). There is no limit to the number of nested groups. The do panel is the same as that of the previous screenshot, whereby the kettle should be turned on and the curtains opened in the bedroom.

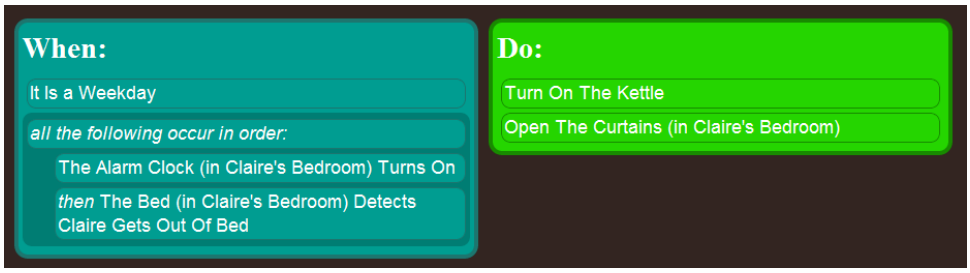


Figure 4: Sample Homeric Policy: Level 2 (intermediate)

Finally, the last of these examples demonstrates a policy which can be written only by those who selected the advanced level. This therefore allows them to have conditional actions within the do panel. As can be seen, the when panel is the same as that of the previous example. The do panel, however, contains a condition that must be true before the kettle is turned on. This condition is simply a check that the kettle has water. The curtains will open regardles of this condition. It is possible to have an else clause paired with the condition, to represent actions which should take place if the condition evaluates to false. As with the nature of the groups in the when panel, the children of the if (and else) are allowed as many actions and conditions as desired.

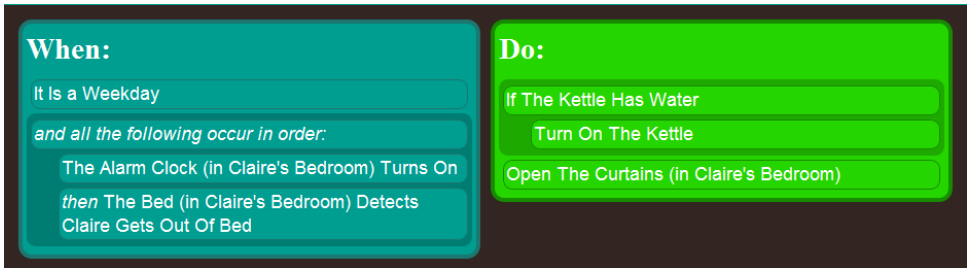


Figure 5: Sample Homeric Policy: Level 3 (advanced)

3 The Survey

3.1 Format

The Google Form service was used to create the online evaluation. The evaluation was designed to take between 15 and 30 minutes to fill in. It was split into five consecutive parts: demographics, parsing policies, transcribing

policies, writing policies and finally concluding questions and remarks. Each of these parts is discussed in turn below.

3.1.1 Part 1: Demographic Questions

The first part of the study included six questions to obtain some basic demographic information about the participant, including gender, age, status and home owner. Additionally, information more specific to technology was also gathered to provide the author with a further understanding of the type of user. Due to the nature of the topic it was important to ensure that results from a good mix of technical backgrounds were obtained. This information both helped ensure data from a representative sample of participants and offered insight into trends based on this information.

At this stage personalised information was requested to customise the Homeric Wizard experience. This included the participant's first name or nickname, and any names of those who they live with. The participants were told that this information would not be stored with their evaluation results. Also included in this personalisation section was the participant's preferred difficulty level.

3.1.2 Part 2: Parsing Policies

Part 2 firstly presented a very simply tutorial describing the when and do panels of the wizard – simply the labelled screenshot shown in Figure 1. The participant was then asked to translate three sample policies which were presented using the Homeric Wizard into their own words. The samples were consistent in theme, but varied in difficultly for each of the three possible difficulty levels. The participant was simply shown the samples at the level they chose in part 1. The first sample described a morning routine, the second an after-work routine, and the third involved heating management. Figures 3, 4 and 5 show examples at the three different difficulty levels.

3.1.3 Part 3: Transcribing Policies

A second-stage tutorial was provided which introduces how to use the wizard, and explains the notion of perspectives at a high level. The participant is then asked to translate two natural language goals into a rule using the Homeric Wizard. This resulted in presenting the Homeric Wizard interface twice, one for each rule. Between these two interfaces, one had perspectives turned on, and the other had them turned off. This was randomised across participants, so some had perspectives turned on for writing goal 1, and off for writing goal 2, and some had this the opposite way round. This was done to ensure there the order of the tasks or the content of the goal did not interfere with the evaluation of perspectives.

The two goals the participant was asked to transcribe were:

- You want the home to turn on your kettle (for a nice cup of tea) when you get up on Tuesdays.
- You are security conscious and during the night you want to make sure your home is secure.

After transcribing each rule the participant was asked how confident they felt that their rule would achieve what they intended. The possible answers were: "Very Unconfident", "Unconfident", "Neutral", "Confident" and "Very Confident" based upon the Likert scale. The participant also had the opportunity to leave any comments.

3.1.4 Part 4: Writing Policies

At this penultimate stage of the evaluation the participant was asked to write any two rules for their home they desired. Just as was done with Part 3, one Homeric Wizard interface was loaded with perspectives, and one without. The participant was also presented with the same confidence questions, as well as the opportunity to leave any comments.

3.1.5 Part 5: Concluding Questions

At this final stage, the participant was presented with the following three questions:

- Using perspectives made it easier to write rules. Possible answers: "Strongly Disagree", "Disagree", "Neutral", "Agree" and "Strongly Agree".

- I found writing the rules within this evalation challenging. Possible answers: “Strongly Disagree”, “Disagree”, “Neutral”, “Agree” and “Strongly Agree”.
- Could you imagine yourself programming your home, using similar style tools as the wizard within this evalution? Possible answers: “Very Unikely”, “Unlikely”, “Neutral”, “Likely”, and “Very Likely” which are based on the Likert scale.

Finally, the participant was provided the opportunity to leave any closing remarks or comments.

3.2 Review

The evaluation was reviewed by experienced academics to ensure the technical correctness of the evaluation format and questions.

3.3 Pilot

A pilot was carried out with three participants who varied in age, technical experience, and familiarity with this field of research to confirm that the evaluation was suitable for the wide range of participants which it was aimed for. The comments were all taken into account and additional tutorials had to be provided within the evaluation form to remove confusion in some parts.

4 The Participants

Since a wide demographic participant set was required it was important to propagate the evaluation to as many different people as possible, and ideally people that the author did not know. This was achieved by using snowball sampling – whereby the author sent the evaluation to friends, family and colleagues and asked them to send it on to others. This resulted in 71 people filling in the evaluation over the duration of two weeks. Results were received for each representative demographic and shown in Tables 1 to 9.

	Male	Female	Totals
Under 21	1	3	4
21 - 30	21	7	28
31 - 40	8	6	14
41 - 50	2	11	13
51 - 60	4	1	5
Over 60	2	5	7
Totals	38	33	71

Table 1: Age and Gender Crosstabulation

	Student	Employed	Unemployed	Retired	Homemaker	Totals
Under 21	4	0	0	0	0	4
21 - 30	14	13	1	0	0	28
31 - 40	2	11	0	0	1	14
41 - 50	0	13	0	0	0	13
51 - 60	0	3	0	2	0	5
Over 60	0	0	0	7	0	7
Totals	20	40	1	9	1	71

Table 2: Age and Status Crosstabulation

	Home Owner		
	No	Yes	Totals
Female	12	21	33
Male	20	18	38
Totals	32	39	71

Table 3: Gender and Home Owner Crosstabulation

	Weak	Competent	Good	Expert	Totals
Female	4	12	12	5	33
Male	0	4	9	25	38
Totals	4	16	21	30	71

Table 4: Gender and Technical Ability Crosstabulation

	Technology Enjoyment		
	No	Yes	Totals
Female	4	29	33
Male	2	36	38
Totals	6	65	71

Table 5: Gender and Enjoyment of Using Technology Crosstabulation

	1 - Easy	2 - Intermediate	3 - Advanced	Totals
Female	14	9	10	33
Male	9	6	23	38
Totals	23	15	33	71

Table 6: Gender and Chosen Level Crosstabulation

	1 - Easy	2 - Intermediate	3 - Advanced	Totals
Poor	4	0	0	4
Competent	9	4	3	16
Good	6	10	5	21
Expert	4	1	25	30
Totals	23	15	33	71

Table 7: Technical Ability and Chosen Level Crosstabulation

	1 - Easy	2 - Intermediate	3 - Advanced	Totals
Under 21	1 (25%)	3 (75%)	0 (0%)	4 (100%)
21 - 30	7 (25%)	5 (18%)	16 (57%)	28 (100%)
31 - 40	3 (21%)	3 (21%)	8 (58%)	14 (100%)
41 - 50	4 (31%)	4 (31%)	5 (38%)	13 (100%)
51 - 60	3 (60%)	0 (0%)	2 (40%)	5 (100%)
Over 60	5 (71%)	0 (0%)	2 (29%)	7 (100%)
Totals	23 (32%)	15 (21%)	33 (47%)	71 (100%)

Table 8: Age and Chosen Level Crosstabulation

	Weak	Competent	Good	Expert	Totals
Under 21	0 (0%)	1 (25%)	3 (75%)	0 (0%)	4 (100%)
21 - 30	0 (0%)	4 (14%)	7 (25%)	17 (61%)	28 (100%)
31 - 40	0 (0%)	2 (14%)	3 (22%)	9 (64%)	14 (100%)
41 - 50	1 (8%)	5 (38%)	4 (31%)	3 (23%)	13 (100%)
51 - 60	1 (20%)	2 (40%)	1 (20%)	1 (20%)	5 (100%)
Over 60	2 (29%)	2 (29%)	3 (42%)	0 (0%)	7 (100%)
Totals	4 (6%)	16 (23%)	21 (29%)	30 (42%)	71 (100%)

Table 9: Age and Technical Ability Crosstabulation

5 Results

The results collected were rich in both quantitative and qualitative data. For the most part this data was kept together to allow any comments made by the participant for any particular question to be taken into consideration when examining the quantitative aspects.

5.1 Homeric

5.1.1 Can Users Understand Homeric?

The author predicted that users would be able to correctly understand the functionality of a sample Homeric policy written in the Homeric Wizard. Participants were presented with three sample rules, and asked to describe each of them in their own words. For each difficulty level the rule was slightly more advanced than that of the level below. Additionally, each rule was slightly more challenging than the previous for all difficulty levels. The author designed a marking scheme to help evaluate the correctness of the participant's natural language description of each rule. Each participant, therefore, had three individual percentage values after marking: one for each example.

To help evaluate if users could understand Homeric rules two hypotheses were written and stastistically verified. Each hypothesis is presented in turn. Figure 6 visually represents the data.

Hypothesis 1: Correctness On average over the three examples, it was predicted that, the average "correctness" of the participants descriptions would be greater than 85%, for each difficulty level.

Level 1: Easy

H_0*: μ is equal to 85%.*

H_1*: μ is greater than 85%.*

A one-sample Z test was performed to calculate if there was a significant difference between the hypothesised 85% correctness and the average correctness score of the participants on level 1. The mean value is 96.8%, with a standard deviation of 9.6%. P is less than 0.001, therefore at a 1% level of significance we can reject H_0 in

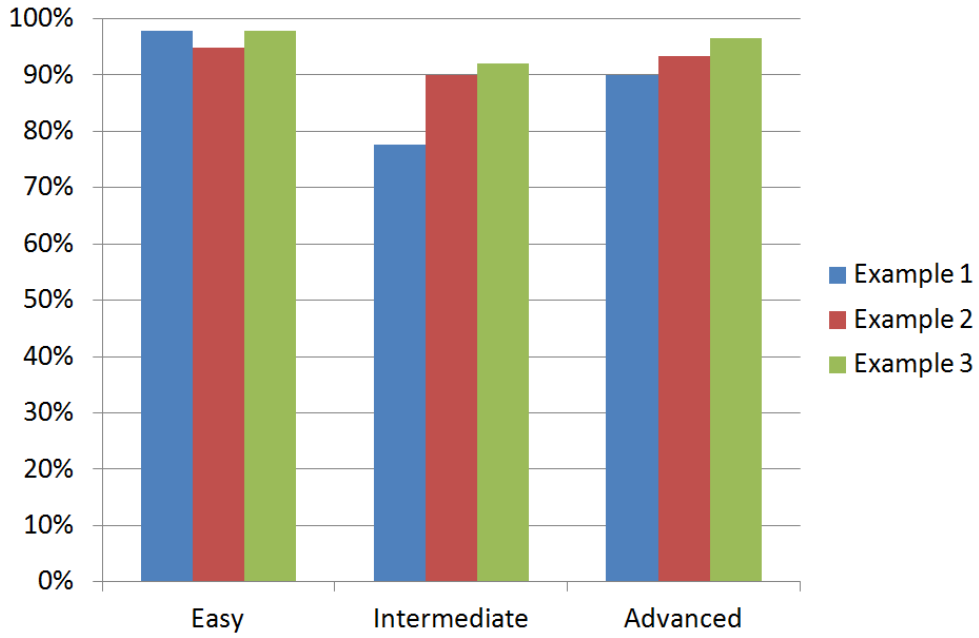


Figure 6: Example Correctness Scores for Each Difficulty Level

favour of H_1 to conclude that μ is greater than 85% and therefore participants using level 1 were succesfully able to describe the three example Homeric rules.

Level 2: Intermediate

H_0*: μ is equal to 85%.*

H_1*: μ is greater than 85%.*

A one-sample Z test was performed to calculate if there was a significant difference between the hypothesised 85% correctness and the average correctness score of the participants on level 2. The mean value is 86.6%, with a standard deviation of 20.5%. P is 0.601, therefore at a 5% level of significance we cannot reject H_0 in favour of H_1. Participants using level 2 were not appearing to understand Homeric policies as succesfully as hoped, however 28 out of the total 45 marks were 100%, and an average of 86.6% is still relatively high.

Level 3: Advanced

H_0*: μ is equal to 85%.*

H_1*: μ is greater than 85%.*

A one-sample Z test was performed to calculate if there was a significant difference between the hypothesised 85% correctness and the average correctness score of the participants on level 3. The mean value is 92.1%, with a standard deviation of 14.7%. P is less than 0.001, therefore at a 1% level of significance we can reject H_0 in favour of H_1 to conclude that μ is greater than 85% and therefore participants using level 3 were succesfully able to describe the three example Homeric rules.

Across All Levels

H_0*: μ is equal to 85%.*

H_1*: μ is greater than 85%.*

A one-sample Z test was performed to calculate if there was a significant difference between the hypothesised 85% correctness and the average correctness score of all the participants. The mean value is 92.5%, with a standard deviation of 1.5%. P is less than 0.001, therefore at a 1% level of significance we can reject H_0 in favour of H_1 to conclude that μ is greater than 85% and therefore participants overall were succesfully able to describe the three example Homeric rules.

Hypothesis 2: Easier Over Time It was hoped that, for across all difficulty levels, the task of describing the rules would become easier for each example, despite the examples increasing in difficulty. This is because the participant would become more familiar with the Homeric Wizard layout of the rules and of the task itself. This can be seen from the data, as the average correctness across all levels for example 1 is 90%, rising to 93% for example 2 and finally 96% for example 3.

Level 1: Easy

H_0*: No significant difference between the correctness score and the example number.*

H_1*: There exists a significant difference between the correctness score and the example number.*

A repeated measures ANOVA test, using a general linear model, was performed to evaluate if there was a significant difference between the correctness score and the example number of the participants on level 1. The mean values for examples 1, 2 and 3 are 97.8%, 94.8% and 97.8% respectively. With an F value of 0.99, df of 2, 44, and a p value of 0.039 we can reject H_0 in favour of H_1 and conclude at a 5% significance level that the means of the correctness scores across the different examples are not equal. However, it can clearly be seen that for those at difficulty level 1 the correctness score decreased for example 2, and remained the same for example 3.

Level 2: Intermediate

H_0*: No significant difference between the correctness score and the example number.*

H_1*: There exists a significant difference between the correctness score and the example number.*

A repeated measures ANOVA test, using a general linear model, was performed to evaluate if there was a significant difference between the correctness score and the example number of the participants on level 2. The mean values for examples 1, 2 and 3 are 77.5%, 90.0% and 92.1% respectively. With an F value of 2.68, df of 2, 28, and a p value of 0.086 we cannot reject H_0 in favour of H_1 at a 5% significance level, meaning that there was no significant difference between the correctness scores and the example number for those at level 2. This could be due to the small sample size for those at intermediate level (only 15 participants chose level 2). However, as the graph in Figure 6 shows, the participants did indeed have an increased correctness score for each example.

Level 3: Advanced

H_0*: No significant difference between the correctness score and the example number.*

H_1*: There exists a significant difference between the correctness score and the example number.*

A repeated measures ANOVA test, using a general linear model, was performed to evaluate if there was a significant difference between the correctness score and the example number of the participants on level 2. The mean values for examples 1, 2 and 3 are 89.9%, 93.3% and 96.5% respectively. With an F value of 3.47, df of 2, 64, and a p value of 0.037 we can reject H_0 in favour of H_1 and conclude at a 5% significance level that the means of the correctness scores across the different examples are not equal.

The Tukey results from the repeated measures ANOVA test, using a general linear model, were used to evaluate what the significant differences were between the correctness scores and the example numbers. With a p value of 0.3748, we can conclude that there is no significant difference between example 1 correctness scores and example 2 correctness scores. Similiarly, with a p value of 0.408 we can conclude that there is no significant difference between example 2 correctness scores and example 3 correctness scores. However, at a 5% confidence level (with p value 0.028) we can conclude that there was significant difference between the example 1 and example 3 correctness scores. Although convential methods have not concluded significant differences at each example increase, the pairwise comparisons between 1 and 2, and 2 and 3, show trends of increase. This is seen from the simultaneous confidence intervals response variables, where the average lower and upper differences from example 1 correctness scores to example 2 is -2.8% to 9.2%, and from example 2 to example 3 is -2.6% to 9.3%. Therefore we can appreciate the increasing trend of correctness scores over the examples at the level 3 difficulty.

Across All Levels

H_0*: No significant difference between the correctness score and the example number.*

H_1*: There exists a significant difference between the correctness score and the example number.*

A repeated measures ANOVA test, using a general linear model, was performed to evaluate if there was a significant difference between the correctness score and the example number across all difficulty levels. The mean values for examples 1, 2 and 3 are 89.9%, 93.1% and 96.0% respectively. With an F value of 4.54, df of 2, 140, and a p value of 0.012 we can reject H_0 in favour of H_1 and conclude at a 5% significance level that the means of the correctness scores across the different examples are not equal.

The Tukey results from the repeated measures ANOVA test, using a general linear model, were used to evaluate what the significant differences were between the correctness scores and the example numbers. With a p value of 0.253, we can conclude that there is no significant difference between example 1 correctness scores and example 2 correctness scores. Similiarly, with a p value of 0.333 we can conclude that there is no significant difference between example 2 correctness scores and example 3 correctness scores. However, at a 1% confidence level (with p value 0.0086) we can conclude that there was significant difference between the example 1 and example 3 correctness scores. Although conventional methods have not concluded significant differences at each example increase, the pairwise comparisons between 1 and 2, and 2 and 3, show trends of increase. This is seen from the simultaneous confidence intervals response variables, where the average lower and upper differences from example 1 correctness scores to example 2 correctness scores is -1.6% to 8.1%, and for example 2 to example 3 is -1.9% to 7.8%. Therefore we can appreciate the increasing trend of correctness scores over the examples at all levels of difficulty.

5.1.2 Can Users Translate into Homeric?

The two goals the participant was asked to transcribe were given in Section 3.1.3. A marking scheme was produced for each goal, just as was done for checking the correctness of the participant answers in the previous question. Marks were allocated exactly the same for all difficulty levels, checking that the basic logic was achieved. The number of events involved in the rules was also calculated, along with the number of language features used by those using level two and three, such as nested logic and conditional actions.

Hypothesis 3: Correctness It was hypothesised that participants would succesfully be able to write rules which meet the given goals. Ignoring any errors in the rules, each rule written for goal 1 and goal 2 was marked based on it containing the events which would result in meeting the goal. Using the same approach as for Hypothesis 1, it was predicted that for each difficulty level the average correctness score across both goals would be greater than 85%. Figure 7 visually represents the data.

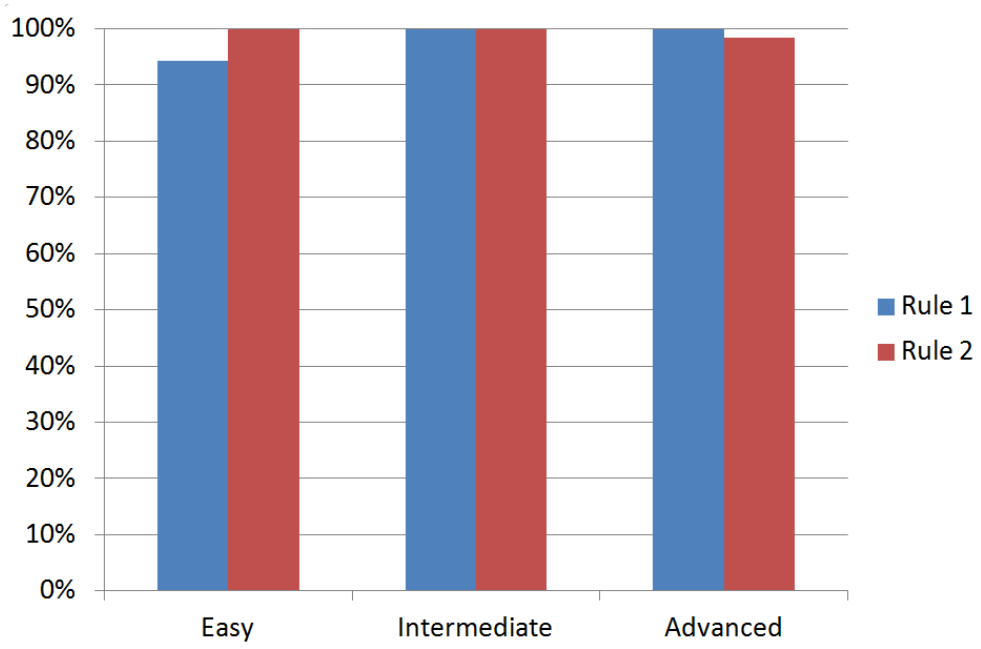


Figure 7: Translation Correctness Scores for Each Difficulty Level

Level 1: Easy
H_0: *μ is equal to 85%.*
H_1: *μ is greater than 85%.*

A one-sample Z test was performed to calculate if there was a significant difference between the hypothesised 85% correctness and the average correctness score of the participants on level 1. The mean value is 97.1%, with a standard deviation of 9.4%. P is less than 0.001, therefore at a 1% level of significance we can reject H_0 in favour of H_1 to conclude that μ is greater than 85% and therefore participants using level 1 were succesfully able to write two rules which met the requirements of the two goals.

Level 2: Intermediate

H_0*: μ is equal to 85%.*

H_1*: μ is greater than 85%.*

A one-sample Z test was performed to calculate if there was a significant difference between the hypothesised 85% correctness and the average correctness score of the participants on level 2. The mean value is 100%, with a standard deviation of 0.0%. P is less than 0.001, therefore at a 1% level of significance we can reject H_0 in favour of H_1 to conclude that μ is greater than 85% and therefore participants using level 2 were succesfully able to write two rules which met the requirements of the two goals.

Level 3: Advanced

H_0*: μ is equal to 85%.*

H_1*: μ is greater than 85%.*

A one-sample Z test was performed to calculate if there was a significant difference between the hypothesised 85% correctness and the average correctness score of the participants on level 3. The mean value is 99.2%, with a standard deviation of 0.8%. P is less than 0.001, therefore at a 1% level of significance we can reject H_0 in favour of H_1 to conclude that μ is greater than 85% and therefore participants using level 3 were succesfully able to write two rules which met the requirements of the two goals.

Across All Levels

H_0*: μ is equal to 85%.*

H_1*: μ is greater than 85%.*

A one-sample Z test was performed to calculate if there was a significant difference between the hypothesised 85% correctness and the average correctness score of the participants on level 3. The mean value is 98.7%, with a standard deviation of 6.9%. P is less than 0.001, therefore at a 1% level of significance we can reject H_0 in favour of H_1 to conclude that μ is greater than 85% and therefore participants overall were succesfully able to write two rules which met the requirements of the two goals.

5.1.3 Can Users Write Homeric?

Participants were asked to write two rules for their home, with no predefined goals. The technical correctness, number of errors and use of language features were all examined, and the results were partnered with the data collected for transcribing two rules previously, to help evaluate if users can succesfully write Homeric.

Hypothesis 4: Very Few Participants Made Errors It was observed that some rules contained logic that would not allow the rule to ever fire (such as "when it is Saturday and it is Sunday"). It most commonly appeared to be assumed that such pairs of statements would be handled as optional rather than both required. In reality, Homer has an integrated rule conflict detector which would observe such situations and be able to report them to the user. However, for this evaluation no such conflict feedback was provided to the user, to ensure that the original rule that the participant felt was complete could be observed.

The author hypothesised that very few participants would write rules which would contain an error (a rule was considered to contain an error if the when clause would never be able to fire due to conditions specified incorrectly). It was hoped that no more than 10% of participants, therefore 7, would write a rule with an error, out of all four rules they had to write independently (the two translation rules from Section 5.1.2, and the two open rules from this section).

Level 1: Easy

H_0*: The data is consistent with a specified distribution of 10% with errors, 90% without.*

H_1*: The data is not consistent with a specified distribution of 10% with errors, 90% without.*

A Chi-Square Goodness-of-Fit test for observed counts was performed to calculate if there was a significant evidence to dispute the data not following the hypothesised 10% of level 1 participants making an error when writing a rule. Table 10 shows the expected count values. The Chi-Square result is 0.044, with a p value of 0.835. Therefore at a 5% level of significance we cannot reject H_0 in favour of H_1, concluding that indeed approximately 90% of participants at level 1 made no errors in their rules.

	Observed	Expected
With Error	2	2.3
No Error	21	20.7

Table 10: Expected Counts for Level 1 Making Errors

Level 2: Intermediate

H_0*: The data is consistent with a specified distribution of 10% with errors, 90% without.*

H_1*: The data is not consistent with a specified distribution of 10% with errors, 90% without.*

A Chi-Square Goodness-of-Fit test for observed counts was performed to calculate if there was a significant evidence to dispute the data not following the hypothesised 10% of level 2 participants making an error when writing a rule. Table 11 shows the expected count values. The Chi-Square result is 1.67, with a p value of 0.2. Therefore at a 5% level of significance we cannot reject H_0 in favour of H_1, concluding that indeed approximately 90% of participants at level 2 made no errors in their rules.

	Observed	Expected
With Error	3	1.5
No Error	12	13.5

Table 11: Expected Counts for Level 2 Making Errors

Level 3: Advanced

H_0*: The data is consistent with a specified distribution of 10% with errors, 90% without.*

H_1*: The data is not consistent with a specified distribution of 10% with errors, 90% without.*

A Chi-Square Goodness-of-Fit test for observed counts was performed to calculate if there was a significant evidence to dispute the data not following the hypothesised 10% of level 3 participants making an error when writing a rule. Table 12 shows the expected count values. The Chi-Square result is 0.03, with a p value of 0.86. Therefore at a 5% level of significance we cannot reject H_0 in favour of H_1, concluding that indeed approximately 90% of participants at level 3 made no errors in their rules.

	Observed	Expected
With Error	3	3.3
No Error	30	29.7

Table 12: Expected Counts for Level 3 Making Errors

Across All Levels

H_0*: The data is consistent with a specified distribution of 10% with errors, 90% without.*

H_1*: The data is not consistent with a specified distribution of 10% with errors, 90% without.*

A Chi-Square Goodness-of-Fit test for observed counts was performed to calculate if there was a significant evidence to dispute the data not following the hypothesised 10% of participants across all levels making an error when writing a rule. Table 13 shows the expected count values. The Chi-Square result is 0.12, with a p value of 0.72. Therefore at a 5% level of significance we cannot reject H_0 in favour of H_1, concluding that indeed approximately 90% of participants across all levels made no errors in their rules.

	Observed	Expected
With Error	8	7.1
No Error	63	63.9

Table 13: Expected Counts for All Level Making Errors

Hypothesis 6: Number of Events Used Increases with Difficulty Level It was hypothesised that as the difficulty level increased, so too would the average number of terms used in the four rules written (the two translation rules from Section 5.1.2, and the two open rules from this section).

H_0*: No significant difference between the number of terms used and the difficulty level.*

H_1*: There exists a significant difference between the number of terms used and the difficulty level.*

A repeated measures ANOVA test, using a general linear model, was performed to evaluate if there was a significant difference between the number of terms used in a rule and the difficulty level of the participant. The mean number of terms in a rule for difficult level 1, 2 and 3 are 3.37, 5.68 and 6.09 respectively. With an F value of 15.38, df of 2, 278, and a p value < 0.001 we can reject H_0 in favour of H_1 and conclude at a 1% confidence level that there exists a significant difference between the number of terms used and the difficulty level.

The Tukey results from the repeated measures ANOVA test, using a general linear model, were used to evaluate what the significant differences were between the number of terms used and the difficulty level. With a p value of 0.0005, we can conclude that at a 1% confidence level there is a significant difference between level 1 average number of terms and level 2 average number of terms. Similarly, with a p value < 0.001 we can conclude at a 1% confidence level that there is a significant difference between level 1 average number of terms and level 3 average number of terms. However, with p value of 0.761 there is no significant difference between the average number of terms for level 2 and 3.

To conclude, it can be seen that there is a significant difference between the average number of terms used by a beginner and an intermediate. However, although a small increase can be seen between intermediate and advanced participants, this is insignificant.

Hypothesis 7: Number of Language Features Used Increases with Technical Ability It was believed that as the technical ability of the participant increased, so too would the number of language features used. Language features include all when clause operators (required - "and", optional - "or", ordered - "then"), and conditional actions in the do clause. The Homeric Wizard controlled which levels were able to make use of which language features. Those who selected level 1 could not make use of any language features. Those at level 2 could make use of the when clause operators, and finally those at level 3 could make use of all language features.

		Total And, Or, Then Operators	Total Conditional Actions	Average Features / Rule	Total Participants
Level 1	Weak	-	-	-	4
	Competent	-	-	-	9
	Good	-	-	-	6
	Expert	-	-	-	4
	Totals	-	-	-	**23**
Level 2	Weak				0
	Competent	0	-	0.00	4
	Good	22	-	0.55	10
	Expert	0	-	0.00	1
	Totals	**22**	**-**	**0.14**	**15**
Level 3	Weak				0
	Competent	12	4	1.63	3
	Good	7	18	1.25	5
	Expert	52	24	0.76	25
	Totals	**71**	**46**	**0.91**	**33**
All Levels	Poor	0	0	0	4
	Competent	12	4	0.25	16
	Good	29	18	0.59	21
	Expert	52	24	0.63	30
	Totals	**93**	**46**	**0.37**	**71**

Table 14: Features Used for Each Difficult Level and Technical Ability

As can be seen from Table 14 this is very much not the case at each level. For level 2 those participants who described their technical ability as "good" used 0.55 language features on average, compared with none by "experts". Even more surprisingly, at level 3 a higher number of language features were used by those less technically competent. 1.63 language features used on average by "competent" participants, compared with 1.25 by "good" and 0.76 by "experts".

Overall, however, when taking into consideration all levels an increasing trend can be seen (weak: 0.0, competent: 0.25, good: 0.59, expert: 0.63). The following hypothesis was evaluated.

H_0*: No significant difference between the difficulty level and average number of language features used per rule.*

H_1*: There exists a significant difference between the difficulty level and average number of language features used per rule.*

A repeated measures ANOVA test, using a general linear model, was performed to evaluate if there was a significant difference between the difficulty level and the average number of language features used per rule. With an F value of 4.52, df of 3, 277, and a p value of 0.004 we can reject H_0 in favour of H_1 and conclude at a 5% confidence level that there exists a significant difference between the difficulty level and the average number of language features used per rule.

The Tukey results from the repeated measures ANOVA test, using a general linear model, were used to evaluate what the significant differences were between the difficulty level and the average number of language features used per rule. The only significant difference detected at the 5% confidence level is the relationship between level 1 and level 4, as well as between level 2 and level 4.

To conclude, examining the data for each difficulty level reveals that there is no trend for an increase in the number of language features used as technical ability increases. However, when the data is combined across all difficulty levels there is significant difference between the average number of language features used by "weak" and "competent" participants with "expert" participants. The data also shows that there is a general trend where the number of average language features used increases with technical ability.

5.1.4 Additional Findings

At each stage of the evaluation the participant had the opportunity to write any comments. The comments related to the core rule language are discussed within this section.

Variable Names When asked to describe sample rules in their own words, it was observed that many participants (across all demographics) translated figures, such as time and temperatures, into their own variables. Such examples include: "tea time", "bed time", "dinner time", "work day", "ambient temperature", and "too cold". This demonstrates a feature that Homer currently does not support, but should strongly consider.

Conditional Actions The concept of conditional actions was discussed by some participants. Some liked it and wanted it, whilst others felt that "muddling conditions in the actions column is messy". One participant wrote a rule entirely within the do clause, until realising for himself that this would never "be called". He commented, explaining this and observing that the relationship between the when and the do is interesting, however he felt that this should be more explicit. In reality, a rule would not be accepted by Homer if it did not contain any terms within the when clause, so this is not a problem. One individual on level 3 did not appreciate what conditional actions were for. He noticed that he was allowed to add conditions to his do panel but said this "seems to suggest that a sunrise (for example) is something you can do?".

There were a few participants who were either on a level below advanced, or had not observed the possibility, and requested conditional actions. For example: "conditional sub statement as part of the 'Do' where I could say 'And switch the kettle on, but only if there is water in it'.", and "if statements within the Do: action. i.e. When: any of the following occur(a,b,c,d,e,f,g) Do: if it was (a) do (something specific to a) if it was (a,b) do (something specific to a and b) always: do (something for all of them".

From this evaluation it was observed that conditional actions typically come down to personal preference. They are not required when representing logic, as multiple smaller rules can be written which would achieve the same goal. Conditional actions are purely designed to allow rules to be condensed into fewer rules, and offer flexibility and language niceties for those who desire it. 46 conditional actions were used in total across the evaluation, with 18 out of the 33 participants at level 3 (the only level which supported conditional actions) choosing to use at least one conditional action throughtout the four rules that they wrote. To conclude, conditional actions definitely have their place in Homeric, but should be made an end user application option, allowing users to choose if they would like the possibility to use them or if they would rather such an option was hidden.

Rule Durations An interesting question was raised by a few of the participants: if a series of actions is performed, such as locking doors and closing windows, when do these actions get reversed? Particpants expressed confusion about whether they should somehow be specifiying when their actions are reversed, or cancelled, or if it is okay for this to be done manually (for example, a rule which locks the front door at night, then the door becomes unlocked manually the next day when you leave the house). One participant expressed the desire to: "open the windows for 30 minutes", and another to "turn on the radio for 10 minutes". This too could be achieved by considering the actions which should take place on exit of a rule. There are multiple possible solutions to this problem, including the requirement for a second rule to undo any actions if so desired, allowing actions to be specified when the rule exits (as seen in Tasker [9]), offering an automated exit rule which offers to undo all those actions that have an opposite (for example sending an email does not, but locking a door or closing a curtain does), or allowing pairs of rules to be notionally grouped, where one is an entrance rule and one is an exit rule. This is certainly an interesting problem, and one that should be explored further.

Sensor and Actuator Fusion Sensor fusion is the principle of one high-level trigger or condition being used to encapsulate multiple triggers and conditions, and actuator fusion uses one high level action to cause multiple lower level actions to be performed. Simplistic examples are "when any window opens" and "do close all windows". The author has already acknowledged sensor fusion as future work for Homer, and this concept was confirmed by many participants. Different terminology and descriptions were used to describe this notion, including "meta-option", "group all the sensor condition into a single combined 'movement detected in the home' condition", "would be nice to set up rules that could be used with groups of devices, for example 'turn off any electric blanket that is on in an unoccupied room' ", "be nice to reuse existing rules and save as new, especially if the conditions were quite complex. Then I could write something like 'When Morning Wake Up occurs...' where Morning Wake

Up was a reusable set of conditions. Similarly for actions. I like the idea of a 'Party Time' action...". The quantity and range of suggestions confirms that this is an important area that needs addressed within Homer.

Overloading When Clause It was observed that a number of participants tended to overload the when clause, resulting in rules which could simply never evaluate to true. For example "when it is Saturday and it is Sunday and time is 8am", this clearly can never fire because it is not possible for the day to be both Saturday and Sunday, however a number of participants wrote such logic. This is an interesting problem, because in natural language it is perfectly acceptable to say "on Saturday and Sundays I want my heating to come on at 8am", which ultimately results in rules which cannot fire. In such situations, the Homer conflict detector would report that it cannot be both a Saturday and Sunday and therefore the rule would never fire. Research and evaluation of the user's understanding and handling of such feedback should be performed.

Although most of the overloading was regarding time, there was the occasional incident where a participant would overload the when clause with unnecessary conditions. For example, one participant wanted their doors locked and burglar alarm turned on each night at 2230; this simply requires logic such as "when it is 2230 do lock the front door and lock the back door and turn on the burglar alarm". However, the participant wrote "When it is a weekday and it is a weekend and the back door is unlocked and the front door is unlocked and it is later than 10:30pm" for their when clause. Ignoring the weekday and weekend aspect, which could be pointed out by the conflict detector, the policy is not going to fire if the back door or front door was already locked - resulting in a rule which would not behave how the user had intended. How to handle such situations is unclear, but should be acknowledged.

Conflicts Interestingly, not many participants commented on the potential for conflicting rules. Those who did appeared to be perfectly satisfied with offline conflict detection (which Homer uses) rather than run-time detection which can lead to confusion and misunderstanding why the system is not behaving how the user would expect. For example, one participant said: "That was quite fun. It would be interesting to see how you would summarize all the rules in my house, and highlight any conflicting ones.". Since Homer already supports offline conflict detection, these observed comments are not an issue.

Negation Homeric does not offer negation of terms, due to the ambiguity caused in many cases and natural language not reading correctly (<not> an email is received, <not> movement detected, <not> curtains closed, <not> front door detects Claire?), and in many cases the opposite event can be offered reliably by the component itself (kettle is not on, door is unlocked, heating is off). Only two participants commented on the lack of negation support, with their examples: "if email has not been received saying 'leave heating on' turn heating off", and "When <both go to bed>, but not <guests staying over>". These comments are noted, but most likely Homeric will not add negation support.

Do-When Instead of When-Do It was observed that, when participants described the example rules, some would describe the actions first, before describing when they would happen, effectively using a do-when format instead of the when-do format of Homer. Unfortunately, no trends could be observed regarding the demographics, technical abilities, or example format for when individuals would decide to use the do-when format. One respondent, when writing the rules themselves, explained "I tend to pick the ACTION first then the condition". A do-when approach is entirely natural and could be perfectly supported by Homer. It should be a consideration, however, for the user interface design.

Loops Only one participant requested loop logic, which is currently not supported by Homeric. They wanted to be notified when it is getting late and they should go to bed: "Actually, I'd like a loop here to flash the lights ...". This is a niche request, and could be considered for advanced users in the future.

5.2 Perspectives

This section will explore Homer perspectives in relation to the Homeric Wizard tool, evaluating various research questions and hypotheses.

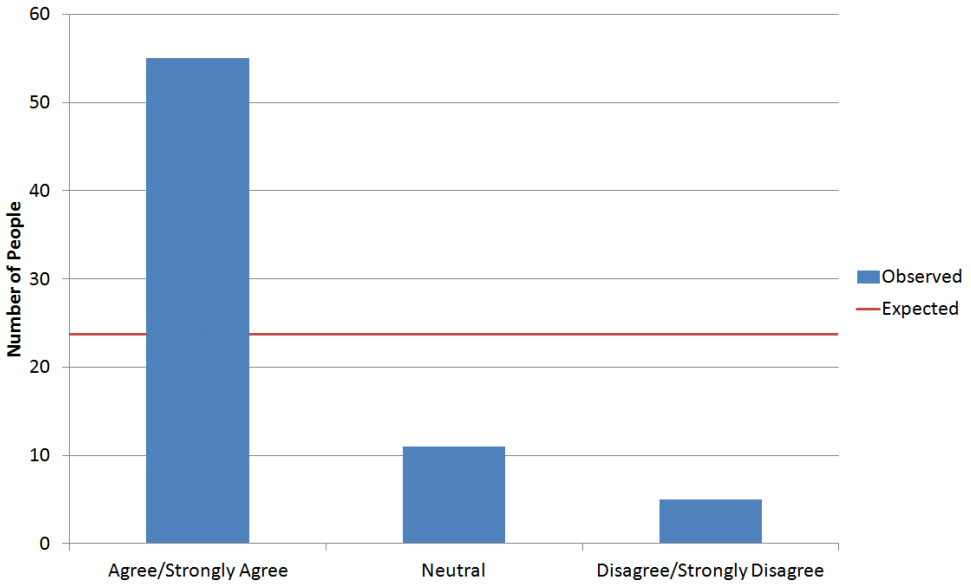


Figure 8: Counts for "Using Perspectives Made it Easier to Write Rules"

5.2.1 Perspectives Ease Writing Rules?

As can be seen from figure 8, participants generally found that using perspectives made writing rules easier, using the Homeric Wizard.

H_0*: The data is consistent with an equally proportionate distribution of participants agreeing/strongly agreeing, neutral and disagreeing/strongly disagreeing.*

H_1*: The data is not consistent with an equally proportionate distribution of participants agreeing/strongly agreeing, neutral and disagreeing/strongly disagreeing.*

A Chi-Square Goodness-of-Fit test for observed counts was performed to calculate if there was significant evidence to dispute the data following a proportionate distribution. The Chi-Square result is 63.0, with a p value < 0.001. Therefore at a 1% level of significance we can reject H_0 in favour of H_1, concluding that indeed the data was not equally proportionate. As the graph in Figure 8 shows, this is because there was a disproportionate number of people agreeing or strongly agreeing that perspectives made writing rules easier.

5.2.2 Respondents Used More Than 1 Perspective?

It was predicted that on average users would use more than one perspective when writing rules using the Homeric Wizard, with perspectives turned on.

H_0*: μ is equal to 2 perspectives on average.*

H_1*: μ is not equal to 2 perspectives on average.*

A one-sample Z test was performed to calculate if there was significant evidence to show that two perspectives were typically used per rule. The mean value is 1.9 perspectives per rule, with a standard deviation of 0.67. P is 0.98, therefore at a 5% level of significance we cannot reject H_0 in favour of H_1 to conclude that μ is equal to 2 perspectives on average and therefore generally participants wrote rules using more than perspective per rule.

5.2.3 Trends Amongst Perspectives and Their Participants?

It was hypothesised that there would be patterns amongst the perspectives used and the demographic or technical background of the participant. However, with only writing two independent rules which use perspectives there was unfortunately not enough data to make statistical conclusions. However, the following tables 15 and 16 show information regarding the perspectives used.

Table 15 shows the percentage of people who made use of each particular perspective at least once in a rule. As can be seen, devices and time are clearly the most common of the four perspectives presented, with locations and people being used by less than 50% of the participants. For each participant the order of the perspectives

displayed to the user in the event library was randomised, to eliminate the presentation of order of perspectives skewing results.

	Devices	Locations	People	Time
% People	93.1%	43.7%	29.6%	90.1%

Table 15: Percentage of People who made use of each Perspective

Table 16 provides information regarding the combinations of perspectives used by each participant. So if a participant used only the devices and time perspectives in their policies, their count would be against the row which has Devices and Time ticked. Interestingly, across all demographics, devices and time remained the most popular combination of perspectives. The author's preconceptions regarding more technically minded individuals preferring devices perspectives, but less technically minded individuals preferring locations or people perspectives, can be disputed by this data.

Devices	Locations	People	Time	Count
✓				1
✓	✓			2
✓	✓	✓		1
✓	✓	✓	✓	5
✓	✓		✓	18
✓		✓		3
✓		✓	✓	10
✓			✓	26
	✓			0
	✓	✓		0
	✓	✓	✓	2
	✓		✓	3
		✓		0
		✓	✓	0
			✓	0
			Total	71

Table 16: Counts of Perspective Combinations Used.

5.2.4 Common Mistakes

The Homeric Wizard interface proved to be rather successful in achieving the goals of evaluating the core Homeric language and perspectives. There were only two types of mistakes across the 71 participants.

The first mistake which was made, rather commonly (21 out of 71 participants – 29.6%), was the understanding of nesting within rule groups. This mistake appears to be made across all demographics and technical backgrounds, resulting in no correlation as to what to type of participant made this mistake. The screenshot in Figure 9 shows an example of this mistake. As can be seen, the natural language text becomes incorrect for having chosen the optional group ("any of the following occur"). Instead of seeing "It is a Saturday or it is a Sunday" as one would expect, participants were presented with "and It is a Saturday and It is a Sunday", which really does not make grammatical nor logical sense. Regardless, the number of people making this mistake, despite being shown three examples of nesting within the examples part of the evaluation form, is far too high and clearly this issue needs addressed.

The second mistake made by only two participants was to read the rules horizontally instead of vertically. Figure 10 shows an example of a rule written by a participant. Although perfectly valid, the participant clearly wrote two separate rules (when the fridge has no milk do play chime notification on phone, and when home temperature falls below 15 degrees turn on the heating) within the same rule statement. Although an issue, it

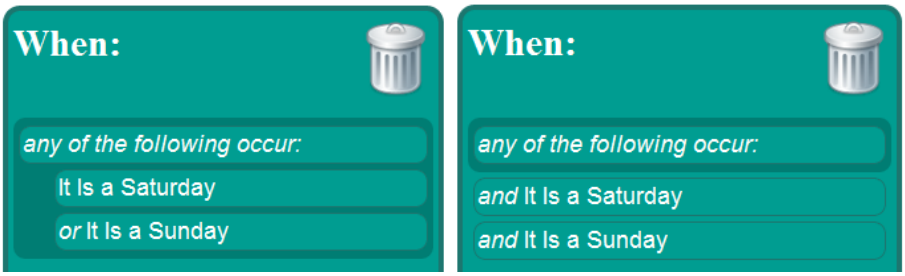


Figure 9: Example of Correct (Left) and Incorrect (right) Nesting in the Homeric Wizard Tool.

would be easily fixed through a more extensive demonstration of the Wizard. It is also the type of mistake that would not be repeated once the participant learnt their mistake.

Figure 10: Example of Incorrect Rule which is written is horizontally

5.2.5 Additional Findings

There were many participants who enthused about perspectives, saying such things as “perspectives make it easier”, “hard to find things without perspectives”, “easier to use when have perspectives”, “I would say perspectives would be an essential part of making a home automation system accessible for most people”, and most simply: “Perspectives = Good”. This feedback, coupled with the quantitative data discussed in previous sections, confirms that for many people perspectives help ease the process of writing rules for the home.

There were only two people who described their dislike for perspectives: “I think I prefer it without perspectives, although I would want to separate time from the list somehow, perhaps in a bit of its own”, and “When I was using the non-perspective view, I like it better but I wanted time to be separate”. Both people, interestingly, noted their desire for time to be separate. This really confirms that they are both individuals who like the devices and time perspectives (given that without perspectives the participant is simply presented with the top level items from the devices perspective). This also matches the results shown in Table 16, where devices and time are the most common group of perspectives to be used.

Two suggestions were made about perspectives. The first of those observed the lack of events within the people category. The participant commented: “Couldn’t find the "Get out of bed at first". Mistakenly looked for it under Person -> Tony -> Events (figuring getting up would be an event of some kind).”. This is a correct assumption, and something that was omitted from the Homeric Wizard tool due to time restrictions in development. However, in reality the people perspective should contain many events to do with people, and ideally events that have been written by the user using sensor and actuator fusion. The second comment that was made regarding perspectives was the suggestion for a fifth perspective: “Another perspective might be "Actions". i.e. I wanted to turn off lots of things. So Action -> Turn off -> List of things I can turn off.”. This does not typically follow the trend of perspectives currently, where the notion is that one is accessing various events in the home from different perspectives. However, it is encouraging that participants were thinking of additional functions and this suggestion should be taken on board.

5.3 Homeric Wizard Tool

5.3.1 Natural Language and Visual Programming Successful?

There were two questions at the end of the evaluation asking if the participant found writing rules for the evaluation challenging, and secondly if they could imagine themselves programming their home using a similar style tool as the Homeric Wizard. These results are presented in this section.

How Challenging? Firstly, it was observed that a large number of people either disagreed or strongly disagreed that they found writing the rules within the evaluation challenging. The figures from strongly disagree to strongly agree are 4.2%, 49.3%, 28.2%, 16.9% and 1.4% respectively. These are shown visually in the graph in Figure 11.

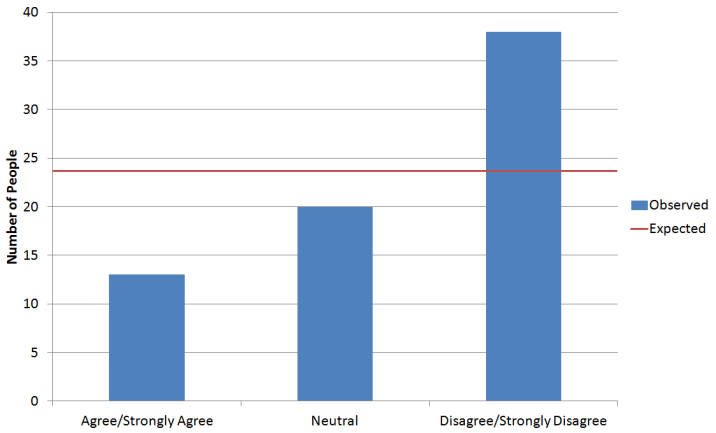


Figure 11: Counts For Participants Agreeing The Evaluation was Challenging.

H_0*: The data is consistent with an equally proportionate distribution of participants agreeing/strongly agreeing, neutral and disagreeing/strongly disagreeing.*

H_1*: The data is not consistent with an equally proportionate distribution of participants agreeing/strongly agreeing, neutral and disagreeing/strongly disagreeing.*

A Chi-Square Goodness-of-Fit test for observed counts was performed to calculate if there was a significant evidence to dispute the data following a proportionate distribution. The Chi-Square result is 14.1, with a p value of 0.001. Therefore at a 5% level of significance we can reject H_0 in favour of H_1, concluding that indeed the data was not equally proportionate. As the graph in Figure 11 shows, this is because there was a disproportionate number of people disagreeing or strongly disagreeing that they found the rules challenging to write. This shows that, across a wide range of demographics and technical abilities, participants were able to formulate rules and, on the whole, not find this exercise too challenging.

It was questioned if age affected how challenging the participant found writing rules. Although there is too little data at each age interval to perform statistical analysis, the graph in Figure 12 visually portrays the slight increase in how challenging the evaluation was to each age category. The graph shows the average response, and the lower and upper quartiles for each age category.

Likely To Use This question in the evaluation unfortunately did not distinguish between those who would not like to program their home at all, and those who would perhaps like to program their home but using a different tool. Nonetheless, Figure 13 shows the spread of counts for those who are likely or very likely to program their home using this tool, those who were neutral to the idea, and those who disagreed or strongly disagreed. As can be seen, there was an extremely strong trend towards those who would be likely (60.6%) or very likely (25.4%) to program their home using the Homeric Wizard tool.

H_0*: The data is consistent with an equally proportionate distribution of participants likely/very likely, neutral and unlikely/very unlikely to program their home using a tool like the Homeric Wizard.*

H_1*: The data is not consistent with an equally proportionate distribution of participants likely/very likely, neutral and unlikely/very unlikely to program their home using a tool like the Homeric Wizard.*

A Chi-Square Goodness-of-Fit test for observed counts was performed to calculate if there was significant evidence to dispute the data following a proportionate distribution. The Chi-Square result is 89.1, with a p value <

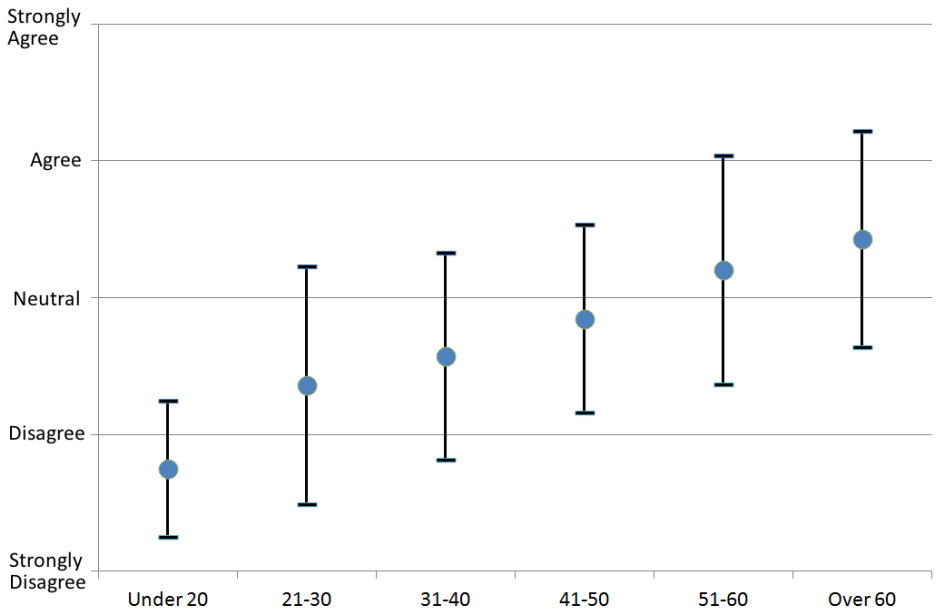


Figure 12: Counts For Participants Agreeing The Evaluation was Challenging at Each Age Bracket.

0.001. Therefore at a 1% level of significance we can reject H_0 in favour of H_1, concluding that indeed the data was not equally proportionate. As the graph in Figure 13 shows, this is because there was a hugely disproportionate number of people selecting likely or very likely to use a tool like the Homeric Wizard to program their home.

5.3.2 Additional Findings

Overall the response and feedback for the Homeric Wizard tool was very positive, as this subset of quotes show: “Looks good and the drag and drop feature works well and is fun to use.”, “panic set in when asked to do the task but found it easy to complete”, “Definitely a fun tool”, “Overall I think it is ace...it’s the first rule based system I have found easyish to get straight away....and perspectives definitely helped!”, "Use of plain language and icons made it simple to set rules", and “Really nice interface! When’s there going to be an iPad version?”.

6 Conclusions

This evaluation has proven extremely valuable in helping understand the success of three main units of work and design. The first of these is Homeric, a custom designed home language to allow rules to be written for the home. The second is the notion of perspectives, which have been designed to ease usability of home systems. Finally, the design of a web-based prototype application for allowing end users to program their home. All three areas were evaluated using an online form, which 71 individuals participated in. This allowed for a large amount of data to be analysed to help evaluate a number of hypotheses regarding each of the three main aspects.

Not only were participants able to translate rules written in Homeric with very little, if any, prior experience or training, they were also able to successfully transcribe and write their own rules for their home using the various language features available. Homeric appeared to meet the requirements of a home language, allowing a wide range of participants from teens to over 70s, and non-technical to highly-technical, to express their own personal goals for their home.

Perspectives prove to ease the rule writing process for many individuals, helping to allow participants of varying technical capabilities to write rules for the home through the same user interface. The four perspectives used were devices, locations, people and time. Devices and time proved to be the most popular perspectives, regardless of demographics or technical experience.

The Homeric Wizard, although just a prototype application to help evaluate design ideas and principles, proved to be a huge success. All participants, no matter their age or experience, were able to successfully write rules using

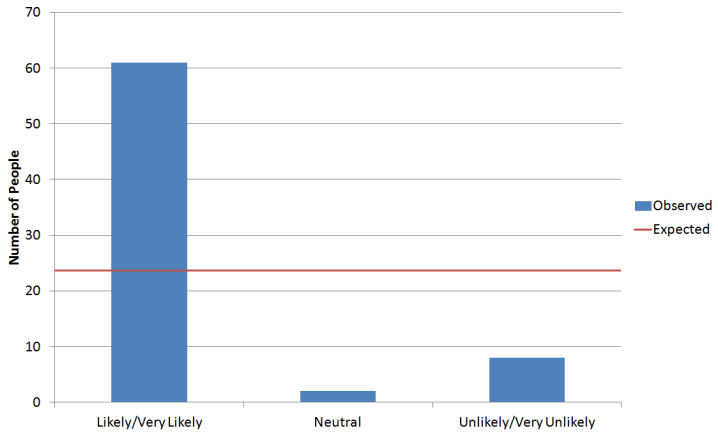


Figure 13: Likelihood of Programming The Home Using the Homeric Wizard.

the tool. Typically it became more challenging as the participants got older, but even those more senior participants were positive and enthusiastic about the tasks: “It was a very interesting and brain exercising exercise!”.

This evaluation, although not its intention, has confirmed how enthusiastic and open minded people are about the notion of programming their homes – from the young: “Hurry up and make it happen!!”, the keen “Personal goals within my home? How long do you have?”, the curious “I enjoyed that and would like to have played about more with different possibilities”, the educated “Rule based systems for "programming" homes is a good idea and perspectives make it much easier”, to the older “Everything is a very good idea, it just makes life more comfortable.” There were also many requests for particular devices, triggers, conditions and actions, showing that participants were able to imagine the possibility of how such a home system could fit into their life and automate their own personal world.

Additionally, there were comments from participants who wanted more than just what was presented to them in the evaluation. They commented on how they would like to add their own custom devices to such a home system, add conditions and actions of their own, and be able to control their home remotely on demand. Homer can offer all of these capabilities, so it is good to hear participants requesting them.

Finally, the challenge of designing and developing a system which allows individuals to easily express rules is immensely difficult, even more so when the users can vary greatly in age and technical experience. Traditionally, the major problem was building a system which could translate the user’s high level, and often ambiguous, goal into concrete logic that a computerised system can understand. The Homeric Wizard tool has made all of this possible. “The challenge for me was working out what rule to set, once I’d decided that, the actual writing part was relatively easy.”

This report has presented a custom policy language, the notion of perspectives and a prototype design for allowing end users to program their home. Each of these have been proven succesful through the evaluation which has been performed, and many directions for future work have been observed.

References

[1] Chris Beckmann and Anind Dey. SiteView: Tangibly Programming Active Environments with Predictive Visualization. In *Adjunct Proceedings of UbiComp*, pages 167 – 168, 2003.

[2] A. F. Blackwell and Rob Hague. AutoHAN: an architecture for programming the home. In *Proceedings IEEE Symposia on Human-Centric Computing Languages and Environments (Cat. No.01TH8587)*, pages 150–157. IEEE, 2001. ISBN 0-7803-7198-4. doi: 10.1109/HCC.2001.995253.

[3] Krzysztof Gajos, Harold Fox, and Howard Shrobe. End User Empowerment in Human Centered Pervasive Computing. In *Pervasive 2002*, pages 134 – 140, 2002. doi: 10.1.1.19.1946.

[4] Claire Maternaghan. *A System to Monitor, Control and Program the Home*. Ph.d. thesis, University of Stirling, Stirling.

[5] Claire Maternaghan. Annual Progress Report: Year 2. Technical Report October, Department of Computing Science and Mathematics, University of Stirling, 2010.

[6] Claire Maternaghan. Homeric Wizard, 2012. URL `www.cs.stir.ac.uk/~cma/homer/homewizard.html`.

[7] Timothy Sohn and Anind K Dey. iCAP: An Informal Tool for Interactive Prototyping of Context-Aware Applications. In *CHI '03 extended abstracts on Human factors in computing systems - CHI '03*, page 974, New York, USA, 2003. ACM Press. ISBN 1581136374. doi: 10.1145/765891.766102.

[8] K. Truong, G. Abowd, and J. Brotherton. Who, what, when, where, how: Design issues of capture and access applications. In *Ubicomp 2001: Ubiquitous Computing*, pages 209 – 224. Springer, 2001. ISBN 978-3-540-42614-1. doi: 10.1007/3-540-45427-6_17.

[9] Erek Zukerman. Tasker For Android: A Mobile App That Caters to Your Every Whim, 2011. URL `www.makeuseof.com/tag/tasker-android-mobile-app-caters-whim/`.