Effective User Interface Designs to Increase Energy-efficient Behavior in a Rasch-based Energy Recommender System

Alain Starke
Eindhoven University of Technology
P.O. Box 513
The Netherlands
a.d.starke@tue.nl

Martijn Willemsen
Eindhoven University of Technology
P.O. Box 513
The Netherlands
m.c.willemsen@tue.nl

Chris Snijders
Eindhoven University of Technology
P.O. Box 513
The Netherlands
c.c.p.snijders@tue.nl

ABSTRACT

People often struggle to find appropriate energy-saving measures to take in the household. Although recommender studies show that tailoring a system's interaction method to the domain knowledge of the user can increase energy savings, they did not actually tailor the conservation advice itself. We present two large user studies in which we support users to make an energy-efficient behavioral change by presenting tailored energy-saving advice. Both systems use a one-dimensional, ordinal Rasch scale, which orders 79 energy-saving measures on their behavioral difficulty and link this to a user's energy-saving ability for tailored advice. We established that recommending Rasch-based advice can reduce a user's effort, increase system support and, in turn, increase choice satisfaction and lead to the adoption of more energy-saving measures. Moreover, follow-up surveys administered four weeks later point out that tailoring advice on its feasibility can support behavioral change.

Keywords

Recommender systems; user experience; energy conservation; behavioral change; Rasch model

1 INTRODUCTION

Although most people acknowledge that household energy conservation is important [36], few actually take effective action [17]. Most governments attempt to change consumers' conservation behavior by educating them [10], typically promoting efficiency solutions as installing Solar PV [17, 34]. However, such approaches have had little impact on individual energy-saving behavior [28, 29].

Research in economics and environmental psychology points out that tailored conservation advice is far more effective in increasing energy-efficient behavior [1, 29, 36, 39, 40]. However, tailored advice usually involves expert advisors visiting households [11, 40], which is hard to scale to larger audiences to have a meaningful impact [2].

Recommender systems could play a major role in moving the energy-saving domain forward. Knijnenburg et al. [24] have shown how they can help consumers to make sense of the different energy-saving possibilities. However, they focused predominantly on tailoring the system's preference elicitation method of the (MAUT-based) recommender to the domain knowledge of the user.

It seems that tailoring conservation advice is challenging, as it remains unclear which energy-saving attributes influence one's decision to adopt an energy-saving measure [1, 6, 12]. In particular, commonly described attributes as execution frequency, cost and impact, as used by Knijnenburg et al. [24], are hard to connect to personal characteristics [12]. Moreover, research has presented mixed evidence about the underlying dimensionality of energy conservation [6, 22, 38].

Starke et al. [35] have shown that these issues could be alleviated by presenting energy-saving advice through the psychometric Rasch model [5, 15], which has two interesting properties for recommender systems. First, persons and energy-saving measures are captured onto a single measurement scale [19], ordered respectively on their energy-saving ability and behavioral difficulty [20, 38]. This order is largely determined by engagement frequencies [5, 15]: measures performed by most people have a low behavioral difficulty (e.g. turning off the lights), whereas those performed by fewer people are more difficult (e.g. installing PV cells) [35, 38]. Conversely, persons engaging in more measures usually have a higher ability, and vice versa. Second, Rasch also connects persons and measures formally. It describes the probability of a person performing a certain measure as a function of that person's ability and the measure's difficulty [5, 19]. For instance, a person whose ability is higher than a measure's difficulty level is likely to perform that measure.

Starke et al. [35] developed a Rasch scale of 79 energy-saving measures and found initial evidence that this scale could be used for recommendations. However, two important questions arise from their research. Firstly, they have not tested how tailored advice would hold up against merely using the Rasch ordering without personalization. The scale's difficulty order could already be sufficient for users to find appropriate measures, rather than also taking the user's ability into account. However, such an approach might lead users to focus on popular, low-difficulty measures [33], which are performed by most users [5]. Hence, presenting measures

RecSys '17, August 27-31, 2017, Como, Italy

ACM ISBN 978-1-4503-4652-8/17/08…$15.00
http://dx.doi.org/10.1145/3109859.3109902

tailored to a user's ability level might be more effective, which we will test and show in study 1.

The second question is what type of advice users should be given? In study 2, we investigate how appropriate tailored recommendations and interface design can help people to improve on their energy savings, exploring the trade-off between offering tailored, but easy and feasible measures, which people might already do, versus tailored but more challenging and novel measures.

We will first discuss in more detail the Rasch model and theories on behavioral change that will bring about the research questions for our two studies. Then, we will discuss the 'Saving Aid', our recommender system that will be used in two different versions. Subsequently, we report two large-scale user studies, which test how ability-based advice should be tailored to optimally help users to increase energy savings through the adoption of more energy-saving measures.

2 THEORY

2.1 The Rasch Model

The Rasch model stems from item response theory (IRT) and assumes persons and measures to share a one-dimensional trait for a certain behavioral goal [5, 20]. This trait manifests itself as a formal measurement scale [19], on which persons and measures are ordered on their ability and behavioral difficulty, respectively [38]. Figure 1 depicts how an energy conservation scale can be considered as a two-sided ruler [20], where measures with a low behavioral difficulty are performed generally more often than difficult ones, and vice versa. On the other hand, persons who perform many measures (e.g. person C in figure 1) are assumed to have a higher ability than those performing few (person A) [38].

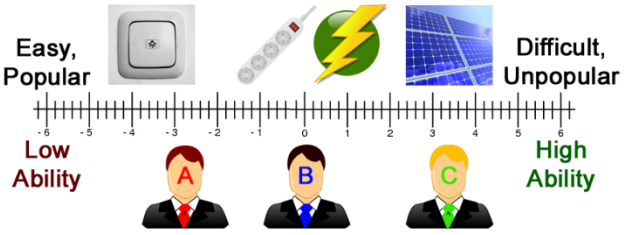

Figure 1: Impression of a Rasch scale of energy-saving measures and persons.

Rasch formalizes the relation between ability and difficulty, describing the probability that a person n performs an energy-saving measure i as the arithmetic difference between that person's energy-saving ability θ and the measure's behavioral difficulty δ [5, 19]:

$$P\{X_{ni} = 1\} = \frac{e^{\theta_n - \delta_i}}{1 + e^{\theta_n - \delta_i}}$$

Figure 2 illustrates for the four energy-saving measures depicted in figure 1 how engagement probability, ability and behavioral difficulty relate to each other. The behavioral difficulty (δ) of a measure is equal to the 50%-engagement probability point of each item-characteristic curve (ICC), which varies between measures. Moreover, the x-axis depicts how different energy-saving abilities (θ), expressed in logistic units or logits, lead to different engagement probabilities (y-axis) for different measures. For instance, a person with energy-saving ability −3 (person A, cf. figure 1) has approximately a 50% probability to regularly turn off the lights (cf. figure 2), but a near zero chance of using green energy or owning solar PV. In contrast, person C (ability = 3.17) has a 50% likelihood of having solar PV installed and is very likely to perform the other three measures as well.

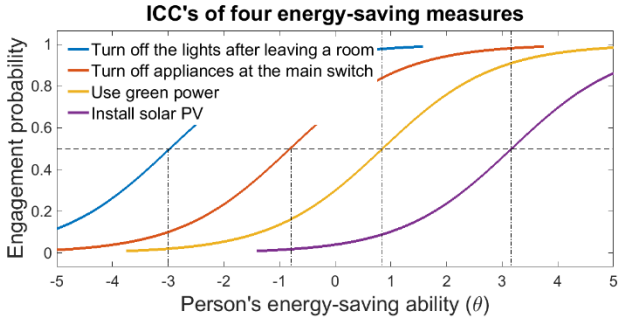

Figure 2: Item-characteristic curves (ICC's) of four energy-saving measures with different behavioral cost levels, yielding different engagement probability levels as a function of a person's energy-saving ability.

The scale's difficulty order can serve as an intuitive tool for those in search of energy-saving possibilities. For example, we could order the measures from very easy (and popular) to very difficult (and infrequently chosen), analogous to showing the most popular items on top of a list, as is commonplace in non-personalized recommender systems [33]. Hence, measures can be considered as a hierarchy of behavioral steps to take towards attaining the goal of saving energy [20], which is an effective representation for users of a recommender system due to its transparency [37].

From a recommender perspective, a classical question arises: would such an ordered scale based on the Rasch model be sufficient or would tailored recommendations outperform such a non-personalized baseline? As tailoring is costly [18], because one needs to know a user's ability before being able to provide good conservation advice, it is important to test its effectiveness.

A person's ability will be a good indicator for where to look for appropriate energy-saving measures on such a scale. Considering the order in figure 1, we should not expect a person who only turns off the lights (e.g. person A) to suddenly install solar PV [38], but rather suggest an intermediate option. The Rasch model prescribes that a person whose ability is equal to a measure's behavioral difficulty has a 50% probability of already performing that measure [5]. Recommending such measures might be a good trade-off between novelty and feasibility, as a person would not perform all measures and they would not be out of reach in terms of behavioral difficulty.

We expect persons to be most motivated to adopt measures with a difficulty level similar to their own ability [32, 35]. This

implies that ability-tailored advice might be an effective approach in helping users to attain their energy-saving goals. If a system indicates which measures are the most appropriate rather than users having to browse the scale's order themselves, then these users should experience less effort in looking for measures, thus feeling supported by the system and likewise experience a higher level of choice satisfaction. These expectations lead to our first research question:

RQ1: Using a Rasch scale to promote the adoption of energy-saving measures, is ability-tailored advice more effective and satisfactory, as well as less effortful than a non-personalized ordered set of measures?

2.2 Supporting a User's Conservation Goals

The Rasch model brings a different perspective to how advice can be tailored, which is particularly useful for an energy recommender system. In most recommender domains, systems aim to recommend items that are rather similar to a user's current behavior [9, 14]. However, Ekstrand and Willemsen point out that such an approach is unable to support users who seek behavioral change [14], which is critical for an energy recommender system [24]. Indeed, energy conservation often relies on goals and long-term commitment [2, 8], as well as the need to learn about novel possibilities to change one's current behavior [3, 17], which are not optimally supported by a behavior-based recommender system [14, 23].

Research in persuasive technology has shown how technological interventions can support the user's goals for attaining a better self [7, 14, 16]. For instance in the energy domain, various tools, such as feedback, goal-setting and social comparison, have successfully influenced people's conservation behavior [27, 30, 31]. However, most persuasion tools have yet to move their personalization beyond the message (how-to advice) and towards the content (what-to advice) [4, 21]. A Rasch-based recommender system can contribute to one's self-actualization by pointing out a user's current performance and suggesting where to look next [14, 23]. Moreover, it could also rely on the user's ability to track long-term progress.

We observe two opportunities to persuade users to make more energy-efficient choices. First, to tailor the difficulty level of energy-saving advice in such a way that it supports sustainable behavior, for example by tailoring it either just above or below the user's ability. Second, to encourage sustainable choices through personalized persuasion, for instance by explaining that a slightly challenging measure is a good match for a user by explaining how well it fits [37].

These opportunities touch upon the same aspect, namely the trade-off between a measure's feasibility and novelty [32]. The Rasch model has this trade-off built-in, as relatively easy measures are often perceived as feasible or attractive [32, 35], but are also likely to be already performed and thus not perceived as novel. Similarly, measures just above one's ability are more likely to be novel, but also might be too challenging. However, we expect that adding a persuasive attribute to a suggested measure might increase its perceived feasibility and, in turn, the likelihood it is adopted [16]. For instance, using a score to indicate that a relatively difficult measure is appropriate might increase its perceived feasibility.

We explore this trade-off for ability-tailored advice in the context of an energy recommender system. This leads to a second research question:

RQ2: How should the difficulty level of Rasch-based advice be tailored to be novel and effective, as well as satisfactory and feasible, and can this feasibility also be influenced by persuasive attributes, such as a fit score?

3 THE 'SAVING AID' CONSERVATION TOOL

We examined our research questions in two different energy recommender system studies. Figures 4a and 4b depict the different interfaces of our 'Saving Aid' web shop, which each relied on a Rasch scale of 79 energy-saving measures developed by Starke et al. [35]. Both systems presented information about different measures and their savings, and encouraged users to choose measures which they were willing to take in the four weeks following system use.

3.1 General Procedure

Figure 3 depicts the general procedure of our studies, showing the sequential steps that each user takes when using one of our 'Saving Aid' interfaces. As a first step, each system determined users' energy-saving ability by surveying them on their current conservation behavior, based on the methodology of Starke et al. [35]. This short survey consisted of 13 energy-saving measures, sampled from different difficulty sub-sets across the entire Rasch scale, for which users had to indicate whether they performed each measure or not, or whether a measure was not applicable to their housing situation.

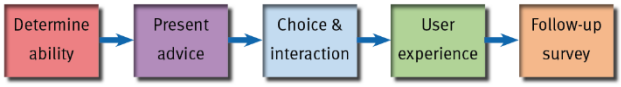

Figure 3: General procedure of our recommender studies. The first four steps were part of our recommender systems, the follow-up survey was sent four weeks later.

Depending on the experimental condition (cf. sections 4.1 & 5.1), advice was presented either in accordance with the user's estimated ability or not. Following this advice, users could interact with the Saving Aid and choose any measure they pleased, as well as review and adjust selected measures in their shopping cart in the upper-right corner of the screen (cf. figure 4a & 4b). Users who confirmed their selection could also opt-in to receive a list of the chosen measures per e-mail.

After finalizing their selection, users were asked to evaluate their system interaction. We examined how users perceived and experienced different interface features through aspects of the framework of Knijnenburg et al. [25, 26]. Perceived support and choice satisfaction were evaluated in both studies, while perceived effort and system satisfaction (study 1), as well as perceived feasibility and novelty (study 2) were study-specific. Finally, four weeks later, we invited all users to a short follow-up survey, in which they were asked to report about the extent to which they performed their chosen measures.

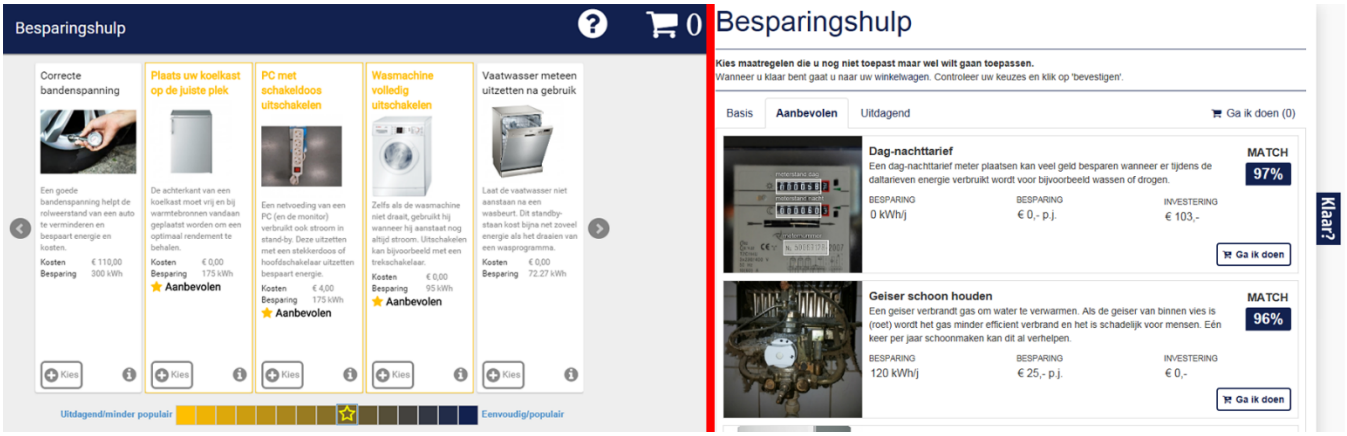

Figure 4a & 4b: The two 'Saving Aid' interfaces employed (Dutch: 'Besparingshulp'), presenting different energy-saving measures. The 'horizontal interface' of study 1 is depicted on the left (4a), in which measures were ordered horizontally, either on ascending or descending order of behavioral difficulty. The 'vertical interface' of study 2 is depicted on the right (4b) and differentiates between three lists of measures: 'basic', 'recommended' (Dutch: 'Aanbevolen'), and 'challenging'. Moreover, in some conditions of the vertical interface, a 'match score' indicated how well a measure fitted the user.

3.2 Unique Interface Aspects

The two systems depicted in Figure 4 have a number of unique interface aspects to examine our research questions. To compare the merits of tailored and non-personalized advice for our first study (cf. RQ1), we employed the 'horizontal interface' shown in figure 4a. The measures were ordered on a horizontal difficulty scale, labeled as ranging from 'easy/popular' to 'challenging/less popular'. Depending on the experimental condition, users were initially presented a particular set of five energy-saving measures, but were free to navigate through all conservation possibilities, either by clicking to see the next measure on the scale or by moving to a different scale position.

In contrast, the 'vertical interface', as shown in figure 4b, addressed RQ2 in study 2. We examined how different levels of ability-based advice and persuasive attributes could lead to more energy-efficient choices. To be able to show a longer list of tailored recommendations, we used a vertical rather than a horizontal interface. The new interface discerned between three lists of energy-saving measures: 'basic', 'recommended' and 'challenging' (cf. figure 4b). Users were initially presented fifteen measures from the 'recommended' list, but were free to navigate both within and between lists. In addition, depending on the experimental condition, users were either shown fit scores or not, indicating how well each measure fitted them.

4 STUDY 1: EFFECTIVENESS OF ABILITY-TAILORED ADVICE

4.1 Research Design

Study 1 investigated whether ability-based advice was more effective and less effortful than a non-personalized approach. Each user of the 'horizontal system' (cf. figure 4a) was assigned to one of four conditions, which determined which five energy-saving measures were initially presented to the user. In this 2x2-research design, we varied the scale's difficulty order (either ascending or descending) and the position on the scale from which measures were initially presented, either tailored (closest to the user's ability) or non-personalized (at the start of the scale, either the most popular or most challenging items). For tailored conditions, three of the five measures were labeled as recommended using a yellow highlight and a star.

4.2 Procedure and Participants

We invited participants from the ThesisTools database to use the Saving Aid depicted in figure 4a. Participants were told they had the opportunity to use a new conservation tool, from which they could choose measures they would like to perform in the four weeks following the study, sent to them per e-mail.

After estimating a user's energy-saving ability (cf. figure 3), the system initially presented five measures, either in accordance with the user's ability or a baseline set. Users could navigate and choose any energy-saving measures they pleased, until confirming their final selection. Subsequently, they were presented a survey on their perceived user experience.

In total, 222 participants completed our web-based experiment and entered a raffle for five 20-euro gift cards. After initial inspection, we excluded 13 participants from analysis, as they did not seem to understand our interface, either choosing all measures which they already performed or making multiple extremely improbable choices from a Rasch model point of view. The remaining sample consisted of 209 participants with a mean age of 44.5 years (SD = 15.8), who chose on average 9.3 energy-saving measures (SD = 8.6), and of which the majority owned a house (63.6%).

Each user was also invited to our follow-up survey, sent approximately four weeks after using the Saving Aid. We sent out invitations to 202 users who had disclosed valid e-mail addresses, from which 78 users (38.6%) replied and indicated to what extent they actually performed their chosen measures.

4.3 Measures

4.3.1 Objective metrics. Several objective metrics could be extracted from the experiment. We derived the users' energy-

saving ability from the initial survey of 13 self-reported items. In addition, to measure the choice behavior of our users, we kept track of outcome variables, such as the total number of chosen conservation measures (log-transformed when included in our final model), the average Rasch difficulty level of these measures and the percentage of those measures being reported as performed in our follow-up survey. From the log data of the interface interactions, we calculated the number of chosen items per navigation click as a measure of objective effort.

4.3.2 Subjective measures. We surveyed our users on four subjective constructs: perceived support, perceived effort, system satisfaction and choice satisfaction. For all survey items, users had to indicate on 7-point Likert scales to what extent they agreed with them, which we submitted to a confirmatory factor analysis (CFA) using ordinal dependent variables. Table 1 only reports three user experience constructs, as we could not distinguish system satisfaction from perceived support due to high cross-loadings. The remaining constructs met the guidelines for convergence validity (cf. table 1), as the average variance explained (AVE) was higher than 0.5 [25], but the internal consistency of the perceived effort construct was somewhat questionable, as Cronbach's alpha for this construct was merely 0.69 [15].

Table 1: Survey items in study 1, with factor loadings and robustness scores per user experience construct. Items without loading were excluded from the CFA and SEM.

PERCEIVED EFFORT – survey items *AVE = 0.511; ALPHA = 0.69*	**Factor Loading**
It took me little effort to use the Saving Aid.	
The Saving Aid takes up a lot of time.	0.804
I quickly understood the functionalities of the Saving Aid.	−0.554
Many actions were required to use the Saving Aid properly.	0.741
The Saving Aid is easy to use.	
PERCEIVED SUPPORT – survey items *AVE = 0.615; ALPHA = 0.81*	**Factor Loading**
I make better choices using the Saving Aid tool.	0.551
The Saving Aid is helpful to find appropriate measures.	0.608
The Saving Aid does not help to come to a decision.	
The Saving Aid presents the measures in a convenient way.	
Because of the Saving Aid, I could easily choose measures.	0.678
CHOICE SATISFACTION – survey items *AVE = 0.598; ALPHA = 0.78*	**Factor Loading**
I am happy with the measures I've chosen.	0.574
I think I've chosen the best measures from the list.	
I would have liked to choose different measures than the ones I've chosen.	
It would be fun to perform the chosen measures.	0.550
The measures I've chosen fit me seamlessly.	0.549

4.4 Results

The objective and subjective constructs, as well as relevant interactions were organized into a path model using Structural Equation Modeling (SEM). Figure 5 depicts the model, which fitted adequately[1]: $\chi^2(108) = 191.000$, $p < 0.001$, *CFI* = 0.957, *TLI* = 0.949, *RMSEA* = 0.061, *90%-CI*: [0.046,0.084].

4.4.1 Tailored advice. We examined whether tailored advice would result in more effective and satisfactory recommendations, as was proposed in RQ1. Figure 5 confirms that users receiving tailored advice, compared to non-personalized suggestions, perceived less effort in system use (coef. −0.440, $p < 0.05$). In turn, perceiving less effort also increased perceived system support (coef. −0.767, $p < 0.001$), which also increased the user's satisfaction with the chosen measures (coef. 0.746, $p < 0.001$).

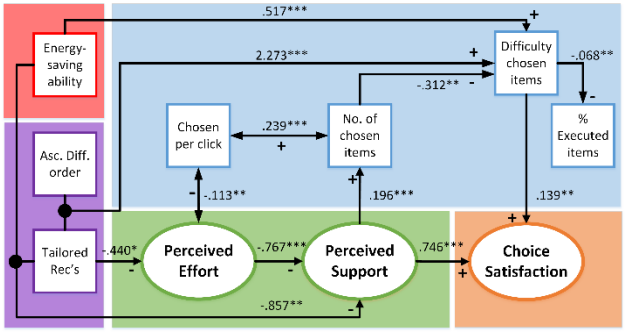

Figure 5: Structural equation model (SEM) for study 1. Numbers on the arrows represent the coefficients. Effects between the subjective constructs are standardized, thus can be considered as correlations. Aspects are grouped by color: Objective system aspects are purple, subjective aspects are green and experience constructs are orange. Behavioral indicators are blue, personal characteristics red [25, 26]. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Besides main effects, figure 5 also depicts that tailored advice interacts with energy-saving ability and the scale's difficulty order. We found high-ability users to perceive less support from tailored advice compared to low-ability users (coef. −0.857, $p < 0.01$), possibly because they were more proficient in navigating through the different energy-saving possibilities. Moreover, users facing a scale in ascending order of difficulty chose measures with a higher level of difficulty, when presented tailored advice (coef. 2.273, $p < 0.001$).

This interaction is easier to interpret when inspecting the marginal effects in figure 6. For non-tailored recommendations, the chosen difficulty was higher when facing a scale in descending difficulty order rather than in ascending order (cf. figure 6a), as users were initially recommended difficult measures. Interestingly, this effect almost reversed for tailored advice, which was arguably due to users having a natural tendency to move rightwards on the scale, starting from the recommendations. This implied for a descending difficulty order that users inspected and chose easier measures relative to those recommended, and vice versa for the tailored ascending condition. As choice satisfaction is positively influenced by

[1] In fitting the model, some issues arose with the discriminant validity of perceived effort [25], as it correlated strongly with perceived support and this correlation was slightly higher than the square root of its own explained item variance (AVE). In other words, perceived effort and support are highly correlated constructs, but they were kept in our model as it helped us understand and explain our results better.

both the difficulty level of chosen items and perceived support (cf. figure 5), the marginal effects in figure 6b show an interesting interaction effect. Although the ascending tailored condition was perceived as more satisfying (compared to non-tailored) due to both difficulty and support, the positive effect of perceived support in the descending tailored condition was counteracted by the negative effect of a lower chosen difficulty.

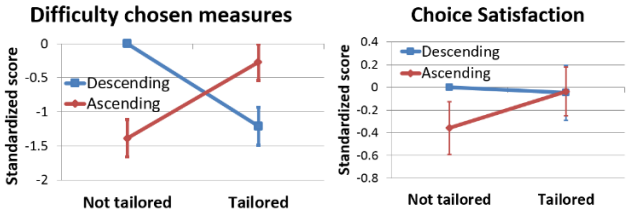

Figure 6a & 6b: Marginal effects of the four experimental conditions on the average difficulty level of chosen measures (6a; on the left), and choice satisfaction (6b; right-hand side). Effects are relative to a baseline of descending & non-tailored. Error bars are 1 std. err.

4.4.2 Mediators: Perceived effort and support. The mediating effort and support constructs depicted in figure 5 had strong behavioral correlates. Effort manifested itself objectively, as users who perceived less effort had also chosen more items for the number of navigation clicks they spent (coef. −0.113, $p < 0.01$). Perceived support also positively affected interaction metrics, as users who perceived higher levels of support also chose more energy-saving measures (coef. 0.196, $p < 0.001$), adhering to earlier findings that a positive user experience can drive energy savings [24]. However, although the difficulty of chosen measures was higher for high-ability users, it decreased if more measures were selected (coef. −0.312, $p < 0.01$).

Figure 5 shows no relation between choice satisfaction and the extent to which users actually performed chosen measures (i.e. '% executed items'). However, we did observe that users mostly followed up on easy measures, as there was a negative relation between the difficulty of chosen system measures and the execution rate four weeks later (coef. −0.068, $p < 0.01$).

4.5 Conclusion

Our path model shows that tailored recommendations positively affect a range of user experience aspects. By reducing both perceived and actual effort, users felt more support, and in turn chose more energy-saving measures and were also satisfied about those choices. This pointed out that providing ability-tailored advice was a more effective approach than merely presenting an ordinal Rasch scale.

The extent to which tailored advice also helps to improve energy savings is less clear from this study. Although more measures were chosen when higher support was perceived, this was against a reduced difficulty level. Moreover, long-term effects in terms of the extent that chosen measures were performed four weeks later, merely showed that users were more likely to perform easier measures.

These findings beg the question whether energy-saving advice should be precisely tailored to a user's ability, or that slightly easier or more difficult advice would be more effective. Hence, presenting tailored, yet slightly more feasible measures might lead to more energy-efficient behavior in the long run. However, users who had chosen relatively difficult measures reported higher levels of choice satisfaction, which could also drive energy savings [24]. Therefore, study 2 examines RQ2, exploring the trade-off between feasibility and novelty for ability-tailored advice.

5 STUDY 2: HOW TO TAILOR THE ADVICE?

5.1 Research Design

Study 2 thus investigated how advice should be tailored to be the most effective, satisfactory and feasible, but also examined whether persuasive attributes can influence this feasibility. All users of our 'vertical system' (cf. figure 4b) were assigned to one out of six conditions, subject to a 2x3-research design. On the one hand, measures were presented either accompanied by a fit score or not. On the other hand, users were initially presented fifteen measures from the 'recommended' tab, which contained measures closest to one out of three levels of ability-difficulty difference: either +1, 0, or -1 logit units (i.e. a relatively low, matching or high difficulty level). By means of these three conditions, we explored the tradeoff between feasibility and novelty/relevance of the measures. Hence, if a user's ability was higher than the item difficulty (+1 logit), users were presented relatively easy measures, which were less novel as users were likely to already perform them (i.e. a 75% probability; cf. figure 2). In contrast, the measures presented at an ability-difficulty difference of -1 logit were more difficult, novel, and relevant, since users were less likely to already perform them (i.e. only a 25%-engagement probability).

The shown fit score, if presented, was a function of the condition's ability-difficulty difference. The top item in the recommended tab would get a fit score close to 100%, while this score decreased for lower-ranked items, following the characteristic curve of the Rasch model (cf. figure 2), with a median value of 77% for the 15th item and most values above 60%. Consequently, measures in the other tabs would show lower fit scores, typically starting at 60-70% and decreasing to 0-1% for items with the largest ability-difficulty difference.

5.2 Procedure and Participants

We invited participants from the ThesisTools participant database to use our 'vertical interface' (cf. figure 4b; only those who did not participate in study 1). The procedure was similar to study 1 and followed the sequential steps depicted in figure 2 (cf. section 3.1): ability estimation, presentation of tailored advice, system interaction and selection of measures, and a user experience survey. Once again, users were free to navigate the entire web shop and could also opt-in to receive the chosen measures via e-mail.

In total, 288 participants completed our experiment and entered a raffle for five web-shop gift cards of 20 euro each. The sample comprised 193 female users (67.0%), 187 home-owners (64.9%), and had a mean age of 39.7 years (SD = 15.0). From this sample, 194 users were also invited to participate in a follow-up survey, four weeks after system use. However, only 46 users (23.7%) indicated to what extent they actually performed their chosen measures.

5.3 Measures

We considered objective measures similar to study 1, such as the total number of chosen measures. Regarding the subjective measures, we considered four constructs: perceived feasibility and perceived novelty (of the recommended measures), as well as perceived support and choice satisfaction. In our final model, perceived novelty was dropped as it suffered from high cross-loadings and including it in our path model deteriorated its fit (cf. section 5.4). Table 2 shows that convergence validity of the other aspects was good, as all AVE values were well above 0.5. Moreover, the internal consistency was good, with high values of Cronbach's alpha.

Table 2: Survey items in study 2, with factor loadings and robustness scores per user experience construct. Items without loading were excluded from the CFA and SEM.

PERCEIVED FEASIBILITY – SURVEY ITEMS *AVE = 0.63; ALPHA = 0.83*	**Factor Loading**
The recommended measures are hard to perform.	−0.772
I do not have the possibility to perform the recommended measures.	−0.792
The recommended measures are applicable in my home environment.	0.741
It takes little effort to perform the recommended measures.	0.736
PERCEIVED SUPPORT – SURVEY ITEMS *AVE = 0.775; ALPHA = 0.92*	**Factor Loading**
I make better choices using the saving aid tool.	0.771
The saving aid is helpful to find appropriate measures.	0.918
Because of the saving aid, i could easily choose measures.	0.923
I would like to use the saving aid more often.	0.802
The saving aid is useless.	
CHOICE SATISFACTION – SURVEY ITEMS *AVE = 0.705; ALPHA = 0.84*	**Factor Loading**
I am happy with the measures I've chosen.	0.674
I think I've chosen the best measures from the list.	
It would be fun to perform the chosen measures.	0.764
The measures I've chosen fit me seamlessly.	0.698

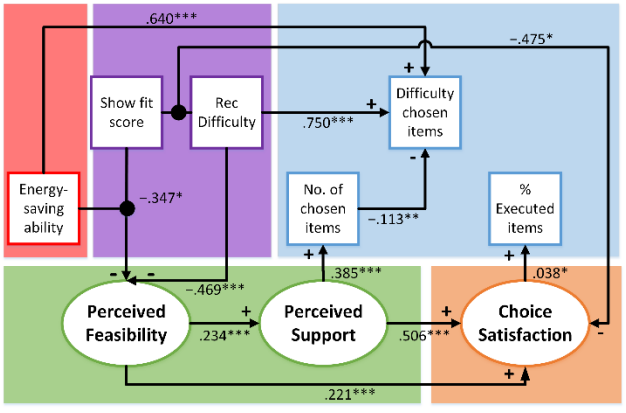

Figure 7: SEM for study 2. Numbers of the arrows represent the coefficients. Effects between the subjective constructs are standardized, thus can be considered as correlations. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

5.4 Results

All variables were organized into a path model using Structural Equation Modeling (SEM). The model had a good fit: $\chi^2(140) = 198.693$ $p < 0.001$, *CFI* = 0.992, *TLI* = 0.990, *RMSEA* = 0.046, *90%CI*: [0.034:0.057]. Figure 7 depicts the fitted path model. We coded the three different difficulty conditions as an ordinal variable in our model (different coding showed similar results), while the fit score was encoded as a dummy variable.

5.4.1 Recommendation Difficulty. We examined whether the difficulty of recommendations affected their perceived feasibility and, in turn, users' satisfaction and choice behavior. We found that comparatively difficult 'recommended' lists were perceived as less feasible (coef. −0.469, $p < 0.001$), confirming that easier measures have a higher perceived feasibility [32].

Figure 7 shows that feasibility also affected other objective and subjective measures. Users who perceived recommended items as feasible also experienced a higher choice satisfaction, both through a direct effect (coef. 0.221, $p < 0.001$), as well as through a mediated effect which increased system support (coef. 0.234, $p < 0.001$) and in turn also choice satisfaction (coef. 0.506, $p < 0.001$). Moreover, similar to study 1, users who perceived more support also chose more items (coef. 0.385, $p < 0.001$). Figure 8a supports these findings, illustrating that a lower recommendation difficulty increased the number of items chosen by a user. Figure 8b shows that a similar effect was also observed for choice satisfaction, which on average decreased if recommendation difficulty increased. However, this effect was mainly observed when a fit score was shown, reflected by the significant interaction effect between fit score and recommendation difficulty (coef. -0.475, $p < 0.05$).

As more satisfied users were more likely to actually execute their chosen items four weeks later (coef. 0.038, $p < -0.05$), it was apparent that presenting comparatively easy tailored recommendations led to a better user interface experience, which in turn led to more energy-efficient choices and sustainable behavioral change.

In contrast with study 1, we observed no relation between the chosen difficulty level and follow-up four weeks later. However, we did observe that presenting relatively difficult measures increased the difficulty level of chosen items (coef. 0.750, $p < 0.001$).

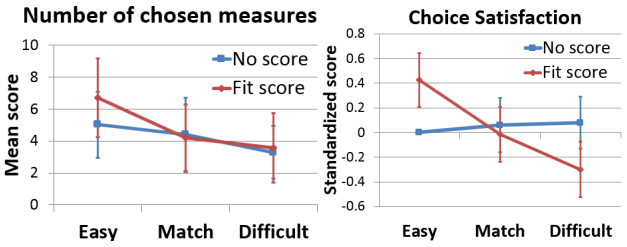

Figure 8a & 8b: The number of chosen measures (8a) and choice satisfaction (8b), as a function of the experimental conditions: fit score (yes or no) and the ability-difficulty difference (easy = +1, match = +0, difficult = -1). Error bars are 1 std. err.

5.4.2 Fit Score. We also examined the effectiveness of presenting a fit score to increase the perceived feasibility, but observed no such main effect. It seems that fit scores reinforced users' preferences for relatively easy measures. Figure 7 shows that high-ability users who were presented a fit score perceived recommended measures as less feasible, as shown by the interaction between these two factors (coef. −0.347, $p < 0.05$). Figure 8b shows that although showing a fit score worsened a user's choice satisfaction for higher levels of recommended difficulty (coef. −0.475, $p < 0.05$), it actually increased choice satisfaction for relatively easy recommendations (compared to showing no fit score). This suggested that fit scores were not successful in motivating users to adopt challenging items, but rather reinforced their current ability levels.

5.5 Conclusion

We investigated how to tailor energy-saving advice to increase its effectiveness and feasibility, and whether this feasibility can be increased by presenting a fit score. We found that ability-tailored recommendations low in difficulty were perceived as more feasible, which made users feel more supported and, in turn, led to more energy-efficient choices and a higher choice satisfaction. In addition, we found that inducing a positive user experience this way, increased the likelihood that users actually performed chosen measures, thus helping them to attain conservation goals, albeit through small behavioral steps in terms of the Rasch scale. This preference for relatively easy measures suggested that users did not mind being presented a recommendation list from which they already performed numerous measures (about 75% of them, according to the Rasch model), compared to a list containing more challenging and novel items with an average engagement probability of 25%. This adhered to earlier research on novelty negatively affecting user satisfaction levels [13].

In addition, we found that presenting a fit score was only effective if the recommended energy-saving measures were relatively easy (cf. figure 7). In fact, showing high fit scores for difficult measures impaired a user's choice satisfaction. Hence, persuasive attributes in our interface were not effective to attain challenging energy-saving goals.

6 DISCUSSION

People often face difficulties in reaching their behavioral goals. To date, most recommender systems are behavior-driven and therefore fail to effectively support those users who wish to accomplish goals for their better selves [14, 23]. In particular, the domain of energy conservation is in need of a tool which helps people to save energy, but research on tailored energy-saving recommendations is limited [24, 35].

We have attempted to fill this gap by using a Rasch scale of energy-saving measures to help users attain their conservation goals [35]. We have established that tailoring energy-saving advice towards a user's ability can reduce a user's perceived effort and increase the feasibility of presented measures. As we also find that users seem to prefer to receive comparatively easy advice which falls just below their own ability [32, 35], we suggest that a user's motivation to save energy is increased the most by making small behavioral steps in terms of the Rasch scale, rather than pushing them to make giant leaps.

This rationale is supported by our findings on the presented fit score. The effectiveness of maximizing fit scores for easy measures might be due to users recognizing presented measures compared to novel suggestions, arguably supporting the credibility of the high fit score and thereby increasing perceived support and satisfaction. In a similar vein, presenting a high fit score for difficult measures might have not been credible and thus deteriorated the interface's efficacy.

The effectiveness of relatively easy ability-tailored advice, especially in terms of feasibility, also suggests that users do not seem to hold particular reasons to not perform certain measures below their own ability level. For example, a person who hardly possesses any electrical appliances might not consider to turn them off at a main switch (cf. figure 1, section 2.1). However, the observed systemic preference for easy measures suggests that this is not the case, and that users might have simply been oblivious to these measures beforehand [3]. However, we do wish to emphasize that ability-tailored advice has shown to be a more effective approach than simply showing the most popular scale items.

Our research is subject to a few limitations. The model fitted in study 1 raises a few concerns regarding the separability of the different constructs. Although this might call for a replication of the study, the outcome of the second study seems to reinforce the findings of the former, as a similar structural model is underlying the results.

Another concern might be that some users might have simply selected a lot of measures and have decided to not perform them. We argue that even though self-report is a delicate measurement method, we have observed small effects of behavioral change four weeks after system use, suggesting that our recommendations and interface have had a positive effect on users' energy-saving behavior.

Finally, the Rasch model has important implications for recommender system research, as well as energy policy. Our Rasch model has shown that a relatively simple personalization algorithm can already yield effective results, which begs the question how an interface using more intricate personalization would perform. Moreover, we have not only measured users' choice behavior, but also checked whether chosen items were actually performed four weeks later. Monitoring a user after system use might be particularly helpful to systems that support behavioral change [14], as it can point out which interface design aspects were effective. However, if a follow-up survey is hard to implement, this research has also shown that system perception constructs can explain a user's choice satisfaction and also correlate with behavioral metrics, which can present a rather complete, triangulated explanation for any effect of system use.

ACKNOWLEDGEMENTS

This work is part of the Research Talent program with project number 406-14-088, which is financed by the Netherlands Organization for Scientific Research (NWO).

We thank Luc Bams, Dennis Dielissen, Tiare Holkema, Lars Middel and Erwin Simons for programming the user interfaces and running the experiments as part of their student projects.

REFERENCES

[1] Abrahamse, W., Steg, L., Vlek, C. and Rothengatter, T. 2005. A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*. 25, 3 (Sep. 2005), 273–291.

[2] Abrahamse, W., Steg, L., Vlek, C. and Rothengatter, T. 2007. The effect of tailored information, goal setting, and tailored feedback on household energy use, energy-related behaviors, and behavioral antecedents. *Journal of Environmental Psychology*. 27, 4 (2007), 265–276.

[3] Benders, R.M.J., Kok, R., Moll, H.C., Wiersma, G. and Noorman, K.J. 2006. New approaches for household energy conservation—In search of personal household energy budgets and energy reduction options. *Energy Policy*. 34, 18 (Dec. 2006), 3612–3622.

[4] Berkovsky, S., Freyne, J. and Oinas-Kukkonen, H. eds. 2012. Influencing Individually: Fusing Personalization and Persuasion. *ACM Trans. Interact. Intell. Syst.* 2, 2 (Jun. 2012), 9:1–9:8.

[5] Bond, T.G. and Fox, C.M. 2006. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Second Edition*. Psychology Press.

[6] Boudet, H.S., Flora, J.A. and Armel, K.C. 2016. Clustering household energy-saving behaviours by behavioural attribute. *Energy Policy*. 92, (May 2016), 444–454.

[7] Cialdini, R.B. 2007. *Influence: the psychology of persuasion*. Collins.

[8] Cialdini, R.B. and Trost, M.R. 1998. Social influence: Social norms, conformity and compliance. *The handbook of social psychology, Vols. 1 and 2 (4th ed.)*. D.T. Gilbert, S.T. Fiske, and G. Lindzey, eds. McGraw-Hill. 151–192.

[9] Cosley, D., Lam, S.K., Albert, I., Konstan, J.A. and Riedl, J. 2003. Is Seeing Believing?: How Recommender System Interfaces Affect Users' Opinions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2003), 585–592.

[10] Costanzo, M., Archer, D., Aronson, E. and Pettigrew, T. 1986. Energy conservation behavior: The difficult path from information to action. *American Psychologist*. 41, 5 (1986), 521–528.

[11] Darby, S. 1999. Energy advice–what is it worth. *Proceedings, European Council for an Energy-Efficient Economy Summer Study, paper III*. 5, (1999), 3–5.

[12] Dietz, T., Gardner, G.T., Gilligan, J., Stern, P.C. and Vandenbergh, M.P. 2009. Household actions can provide a behavioral wedge to rapidly reduce US carbon emissions. *Proceedings of the National Academy of Sciences*. 106, 44 (Nov. 2009), 18452–18456.

[13] Ekstrand, M.D., Harper, F.M., Willemsen, M.C. and Konstan, J.A. 2014. User Perception of Differences in Recommender Algorithms. *Proceedings of the 8th ACM Conference on Recommender Systems* (New York, NY, USA, 2014), 161–168.

[14] Ekstrand, M.D. and Willemsen, M.C. 2016. Behaviorism is Not Enough: Better Recommendations Through Listening to Users. *Proceedings of the 10th ACM Conference on Recommender Systems* (New York, NY, USA, 2016), 221–224.

[15] Embretson, S.E. and Reise, S.P. 2000. *Item Response Theory for Psychologists*. Psychology Press.

[16] Fogg, B.J. 2002. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann.

[17] Gardner, G.T. and Stern, P.C. 2008. The short list: The most effective actions US households can take to curb climate change. *Environment: science and policy for sustainable development*. 50, 5 (2008), 12–25.

[18] Harper, F.M., Li, X., Chen, Y. and Konstan, J.A. 2005. An Economic Model of User Rating in an Online Recommender System. *SpringerLink* (Jul. 2005), 307–316.

[19] Kaiser, F.G., Byrka, K. and Hartig, T. 2010. Reviving Campbell's Paradigm for Attitude Research. *Personality and Social Psychology Review*. 14, 4 (Nov. 2010), 351–367.

[20] Kaiser, F.G. and Wilson, M. 2004. Goal-directed conservation behavior: the specific composition of a general performance. *Personality and Individual Differences*. 36, 7 (May 2004), 1531–1544.

[21] Kaptein, M., Markopoulos, P., de Ruyter, B. and Aarts, E. 2015. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*. 77, (May 2015), 38–51.

[22] Karlin, B., Davis, N., Sanguinetti, A., Gamble, K., Kirkby, D. and Stokols, D. 2014. Dimensions of Conservation Exploring Differences Among Energy Behaviors. *Environment and Behavior*. 46, 4 (May 2014), 423–452.

[23] Knijnenburg, B.P., Sivakumar, S. and Wilkinson, D. 2016. Recommender Systems for Self-Actualization. *Proceedings of the 10th ACM Conference on Recommender Systems* (New York, NY, USA, 2016), 11–14.

[24] Knijnenburg, B.P., Willemsen, M. and Broeders, R. 2014. Smart sustainability through system satisfaction: Tailored preference elicitation for energy-saving recommenders. *Twentieth Americas Conference on Information Systems (AMCIS)* (Georgia, 2014).

[25] Knijnenburg, B.P. and Willemsen, M.C. 2015. Evaluating recommender systems with user experiments. *Recommender Systems Handbook*. Springer. 309–352.

[26] Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H. and Newell, C. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*. 22, 4–5 (2012), 441–504.

[27] Mankoff, J., Matthews, D., Fussell, S.R. and Johnson, M. 2007. Leveraging Social Networks To Motivate Individuals to Reduce Their Ecological Footprints. *Proceedings of the 40th Annual Hawaii International Conference on System Sciences* (Washington, DC, USA, 2007), 87–.

[28] McKenzie-Mohr, D. 2000. Fostering sustainable behavior through community-based social marketing. *American Psychologist*. 55, 5 (2000), 531–537.

[29] McMakin, A.H., Malone, E.L. and Lundgren, R.E. 2002. Motivating Residents to Conserve Energy without Financial Incentives. *Environment and Behavior*. 34, 6 (Nov. 2002), 848–863.

[30] Petkov, P., Goswami, S., Köbler, F. and Krcmar, H. 2012. Personalised Eco-feedback As a Design Technique for Motivating Energy Saving Behaviour at Home. *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design* (New York, NY, USA, 2012), 587–596.

[31] Petkov, P., Köbler, F., Foth, M., Medland, R. and Krcmar, H. 2011. Engaging energy saving through motivation-specific social comparison. *CHI'11 Extended Abstracts on Human Factors in Computing Systems* (2011), 1945–1950.

[32] Radha, M., Willemsen, M.C., Boerhof, M. and IJsselsteijn, W.A. 2016. Lifestyle Recommendations for Hypertension Through Rasch-based Feasibility Modeling. *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (New York, NY, USA, 2016), 239–247.

[33] Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S.M., Konstan, J.A. and Riedl, J. 2002. Getting to Know You: Learning New User Preferences in Recommender Systems. *Proceedings of the 7th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2002), 127–134.

[34] Schultz, P.W. 2014. Strategies for promoting proenvironmental behavior: Lots of tools but few instructions. *European Psychologist*. 19, 2 (2014), 107–117.

[35] Starke, A., Willemsen, M.C. and Snijders, C. 2015. Saving Energy in 1-D: Tailoring Energy-saving Advice Using a Rasch-based Energy Recommender System. *Proc. 2nd Int. Workshop on Decision Making and Rec. Sys. (DMRS)* (Bolzano, Italy, 2015), 5–8.

[36] Steg, L. 2008. Promoting household energy conservation. *Energy policy*. 36, 12 (2008), 4449–4453.

[37] Tintarev, N. and Masthoff, J. 2007. A Survey of Explanations in Recommender Systems. *2007 IEEE 23rd International Conference on Data Engineering Workshop* (Apr. 2007), 801–810.

[38] Urban, J. and Ščasnỳ, M. 2014. Structure of Domestic Energy Saving: How Many Dimensions? *Environment and Behavior*. (2014), 13916514547081.

[39] Wilson, C. and Dowlatabadi, H. 2007. Models of Decision Making and Residential Energy Use. *Annual Review of Environment and Resources*. 32, 1 (2007), 169–203.

[40] Winett, R.A., Love, S.Q. and Kidd, C. 1982. Effectiveness of an energy specialist and extension agents in promoting summer energy conservation by home visits. *J. Environ. Syst.;(United States)*. 12, 1 (1982).