Progress in Adaptive Web Surveys: Comparing Three Standard Strategies and Selecting the Best

Thomas M. Prinz(✉), Jan Plötner, Maximilian Croissant, and Anja Vetterlein

Course Evaluation Service, Friedrich Schiller University Jena, Jena, Germany
{Thomas.Prinz,Jan.Ploetner,Maximilian.Croissant,
Anja.Vetterlein}@uni-jena.de

Abstract. Progress indicators inform the participants of web surveys about their state of completion and play a role in motivating participants with a special impact on dropout and answer behaviour. Researchers and practitioners should be aware of this impact and, therefore, should select the right indicator for their surveys with care. In some cases, the calculation of the progress becomes, however, more difficult than expected, especially, in adaptive surveys (with branches). Previous work explains how to compute the progress in such cases based on different prediction strategies, although the quality of prediction of these strategies still varies for different surveys. In this revised paper of a conference paper, we demonstrate the challenges of finding the best strategy for progress computation by presenting a way to select the best strategy via the RMSE measure. We show the application of this method in experimental designs with data from two large real-world surveys and in a simulation study with over $10k$ surveys. The experiments compare three prediction strategies taking into account the minimum, average, and maximum number of items that participants have to answer by the end of the survey. Selecting the mean as strategy is usually a good choice. However, we found that there is no single best strategy for every case, indicating a high dependence on the structure of the survey to produce good predictions.

Keywords: Progress indicator · Web survey · Prediction strategy · Simulation study

1 Introduction

As web information systems, web surveys are an important tool in evaluation research to provide a fast and straight-forward way to collect information from a user. They usually include a variety of questions or statements to be rated by the participants. In order to not overwhelm the user, these questions are often separated into different pages. To show the participant how much of the survey is left, many surveys include progress indicators (PIs).

A. Bozzon et al. (Eds.): WEBIST 2019, LNBIP 399, pp. 148–167, 2020.
https://doi.org/10.1007/978-3-030-61750-9_7

PIs serve not only informative purposes, but are linked to user's motivation to continue the survey thoroughly and answer every question diligently [6]. Typically, the PI displays the progress in percentage between 0 and 100%. There are three main differences between PIs of web surveys and PIs of other usual tasks in software, e.g., for machine learning [14] and database queries [12]. Participants of web surveys 1) have to focus on the task, 2) can influence the PI, and 3) do not necessarily have an interest on the result of the survey [21]. This means that web survey PIs take on a special role of motivating participants to continue and finish the survey. In keeping with this, it has been shown that web survey participants prefer to have a PI to be aware of their real progress [16,21], which also functions as an indication of how much more effort is needed to finish. However, the computation of the progress can be difficult in case of surveys with *adaptivity* (branches). In previous work [19], we propose an equation to compute the progress in adaptive surveys, based on the *number of remaining items* (questions) at each point of time. This number of remaining items depends on a chosen *prediction strategy*. Such a strategy tries to predict the number of remaining items for each page since the participants may take different paths in the survey with different numbers of remaining items. For example, three known prediction strategies are: 1) take the minimum, 2) average, or 3) maximum number of remaining items [10]. However, we suspect that it depends on the structure of the survey which prediction strategy is the best. Furthermore, the comparison of the quality of the strategies seems to be not trivial.

This paper is a revised version of a conference paper [20]. It is extended with additional related work and examples. Furthermore, it provides more evidence showing that none of the three known mentioned prediction strategies is always the best, based on a simulation study with over $10k$ surveys. One main goal of this research is to find a measure in order to select the strategy that provides a prediction of the remaining items closest to the *true* number of remaining items and, therefore, the *true* progress. The *true* progress is the actual degree of completion of the survey. We support the idea of displaying the *true* progress since research in Human-Computer Interaction (HCI) reveals probable side-effects of PIs on the answer and dropout behaviour of participants [21]. Especially the progress speed (the rate in which the PI increases) seems to influence the decision whether a participant finishes a long survey [17,21]. For example, a PI that is slow at the beginning and gets faster towards the end seems to discourage participants and causes higher dropout rates [1,2,15]. A meta-analysis of PI speeds by Villar et al. [21] supports these observations.

A different pattern can be observed for fast-to-slow PIs, which actually have been found to encourage the participants to finish the survey. While this might seem like a desirable outcome, there is no research regarding the perception of the whole process, i.e., how frustrating the survey was, especially towards the end. The perception of being misguided could therefore decrease any willingness to participate in future surveys and even result in a changing motivation to properly answer the survey throughout.

On the contrary, PIs which try to display the true progress and can therefore be viewed as honest representations should reduce any side-effects, could strengthen long-term motivation, and hold a greater informational value.

The recognition of imprecise PIs by the participants seems to lead to higher dropout rates as a study of Crawford et al. implies [3]. If the growth of the PI is unexpected, e.g., it does not match the amount of time mentioned in the introduction of a survey, the dropout rate increases significantly [6,22].

There are lot of discussions regarding PIs' impact on the dropout behaviour of participants. Some studies suggest that surveys without any PI have lower dropouts than surveys with a PI [3,13]. However, conversely, Heerwegh and Loosveldt state that PIs can have a positive effect on the dropout rate, as they are a highly requested tool to constantly reassess the cost of a given survey [6]. As an indication of when a survey ends could increase the motivation to finish and may therefore reduce the dropout rate [5,6]. When used as a psychological frame of reference, the effects of PIs may differ individually, as the demographic background of the participants seems to decide whether a PI has a positive or negative effect on the completion rate [3]. Other studies on the other hand claim that the effect of the PI on the dropout rate might be negligible altogether [13,22].

Besides the effect on the completion rate, survey design principles and the participant's volition have to be taken into account. As mentioned, most participants prefer to have a PI [6,16,21]. In addition, design principles, e.g., of Dillman et al. [4], claim that PIs should be typical parts of web surveys. In other words, sometimes PIs are necessary to accommodate the participants wishes and there is no room for discussions if a PI should be used or not. In those cases, the effect of the PI on the dropout rate might be small, but could be seen as a mere addition to the positive effects from a perspective of design and preference. As discussed before, this should be the case if the PI tries to show the *true* progress.

If the *true* progress should be displayed, different prediction strategies should be comparable to select the one whose predictions are nearest to the true progress. To reach the aim of a more precise progress computation, we argue in this paper that the *Root Mean Square Error (RMSE)* is the most fitting measure to describe the quality of prediction strategies for progress computation. Researchers conducting surveys can use the RMSE to determine the best known strategy for each survey and can give the participant a PI, which represents the true progress as well as possible. Furthermore, this paper shows as a second goal that the trivial prediction strategies, mentioned earlier, can lead to bad predictions in specific cases and that there is no single best strategy for all surveys. Further research should find solutions for these cases.

The paper has the following structure: First of all, we will explain in Sect. 2 how the computation of progress in adaptive surveys work and how the prediction strategies can be applied. Following this, four different measures as indicators for quality of the prediction strategy will be compared in Sect. 3. Findings will then be applied in experimental designs in Sect. 4. Section 4 argues further which measure is most suitable and explains some disadvantages with current prediction strategies. It also contains a simulation study with over $10k$ surveys. Section 5 provides some concluding thoughts and prospects for further research.

2 Related Work and Preliminaries

Many studies in research take into account the differences in PI speeds. However, it is not the focus of research *how* to calculate an exact progress in web surveys (especially in surveys with high adaptivity). The first work (to the best of the authors' knowledge) is the thesis of Kaczmirek [10] that presents an equation of progress calculation in adaptive surveys. Based on this equation, our previous work [19] describes a general algorithm to predict the number of remaining items, which is part of that equation. The number of remaining items is typically unknown because of the unknown "*path*" in a survey taken by a participant. Our general algorithm allows to apply different prediction strategies to tune the progress as closely as possible to the *true* progress.

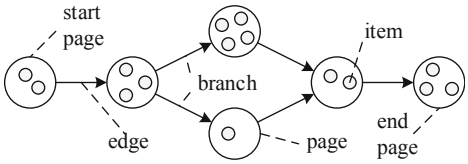

Fig. 1. A simple questionnaire graph (taken from the conference paper [20]).

Surveys can be considered as acyclic, connected and directed graphs (digraphs) where vertices describe pages and edges describe the control flow between these pages. Our above mentioned algorithm is based on such an abstract survey model that is called the *questionnaire graph* or in short *Q-graph*. It is an acyclic, connected digraph $Q = (\mathbb{P}, \mathbb{E})$ with a set of pages $\mathbb{P} = \mathbb{P}(Q)$ and a set of *edges* $\mathbb{E} = \mathbb{E}(Q)$, which connect the pages. Q-graphs have exactly one page without any incoming edge (the *starting page*) and exactly one page without any outgoing edge (the *ending page*). Each page of the Q-graph contains *items* (questions, etc. at the page). Therefore, each page is a finite set $\{i_1, i_2, \ldots\}$ of items $i_1, i_2, \ldots$ which are not specified in detail and are assumed to be unique in this context. Since a page P is a set, $|P|$ is the number of items at P. Figure 1 illustrates a simple Q-graph.

A participant can reach a certain page in the Q-graph if there is a *path* from the current page to that page. A path is a sequence $W = (P_0, \ldots, P_m)$, $m \geq 0$, of pages, $P_0, \ldots, P_m \in \mathbb{P}(Q)$, where an edge exists for each two pages appearing consecutively: $\forall\, 0 \leq i < m \colon (P_i, P_{i+1}) \in \mathbb{E}(Q)$.

In our previous work [19], we generalized Kaczmireks equation to compute the progress in item precision for arbitrary Q-graphs. This equation is recursive and returns values between 0 and 1 (i.e., 0 and 100%):

$$\rho(P) = \rho(P_{prev}) + |P| \frac{1 - \rho(P_{prev})}{rem(P)} \tag{1}$$

It describes how to calculate the progress $\rho(P)$ at the current page P. The calculation of the current progress sums the progress $\rho(P_{prev})$ of the previous

page P_{prev} and the impact on the progress of the current page, $|P|\frac{1-\rho(P_{prev})}{rem(P)}$. If the current page P is the starting page, then the progress $\rho(P_{prev})$ of the previous page is 0. The impact on progress of the current page depends on the number of items $|P|$ at P and the impact of a single item $\frac{1-\rho(P_{prev})}{rem(P)}$. The impact of a single item contains the *remaining progress* $(1-\rho(P_{prev}))$ and the *number of remaining items* $(rem(P))$. The usage of the remaining progress in the equation allows the progress to adopt to the number of remaining items. For example, if a participant follows a branch, which reduces the number of remaining items, then the impact of each item increases, accelerating the growth of the PI. Otherwise, if the number of remaining items increases, the impact with each item decreases, decelerating the growth of the PI.

Input: A Q-graph Q and a selection operator $\sqcup$.
Output: For each $P \in \mathbb{P}(Q)$ the remaining items $rem(P)$.

```
Set rem(P) = 0 for each P ∈ ℙ(Q)
worklist ← queue(ℙ(Q)), visited ← ∅
while worklist ≠ ∅ do
    P ← dequeue(worklist)
    directSucc ← {succ: (P, Succ) ∈ 𝔼(Q)}
    if directSucc ⊆ visited then
        if directSucc = ∅ then
            rem(P) ← |P|
        else if |directSucc| = |{Succ}| = 1 then
            rem(P) ← |P| + rem(Succ)
        else
            rem(P) ← |P| + ⨆_{Succ∈directSucc} rem(Succ)
        visited ← visited ∪ {P}
    else
        enqueue(worklist, P)
```

Fig. 2. The general algorithm for computing the number of remaining items for arbitrary prediction strategies (taken from previous work [19]).

The number of remaining items $rem(P)$ is the only unknown part of (1). This number highly depends on the path a participant takes throughout the survey—which is usually unknown too. Therefore, it is necessary to predict the number of remaining items.

Different prediction strategies are possible making the computation of the progress a challenge. Our general algorithm for calculating the number of remaining items [19] allows such different strategies. An input of the algorithm is a *selection operator* $\sqcup$ representing these strategies. This operator combines different numbers of remaining items to a single prediction if the survey forks. Figure 2 shows the algorithm.

Our algorithm considers exactly three situations during the prediction of the remaining items for a page P in the inner if-then-else-structure: either P has 1)

no successor, 2) exactly one direct successor, or 3) more than one direct successor. The number of remaining items for the first situation 1) is simply the number of items on P, $|P|$, since P has no successor and, therefore, is the ending page. In situation 2), the number of remaining items is the sum of the number of items on P, $|P|$, and the number of remaining items $rem(Succ)$ of its direct successor $Succ$. Different numbers of remaining items may be possible after P in situation 3) since P has multiple direct successors $Succ_1, \ldots, Succ_n$, $n \geq 2$. This situation solves the selection operator that combines all those numbers to a single prediction. So the operator receives all those numbers as input, $\sqcup\big(rem(Succ_1), \ldots, rem(Succ_n)\big)$, and gives a prediction $rem(P)$.

Typical examples of prediction strategies (i.e., selection operators) are the *minimum*, *mean*, and *maximum* functions. Taking the minimum, the number of remaining items is the *smallest* number of items. As a result, the progress is fast at the beginning and becomes slower if the participant takes a path containing more items than the operator has detected. For the maximum, it is vice versa. It represents the *largest* number of items. In the case of the *mean*, the number of remaining items of a page is always the average of numbers of remaining items of the direct successor pages.

Take a look at Fig. 3. It shows a Q-graph with 13 pages where the names of the pages are assigned to the circles. The best solution to traverse the path

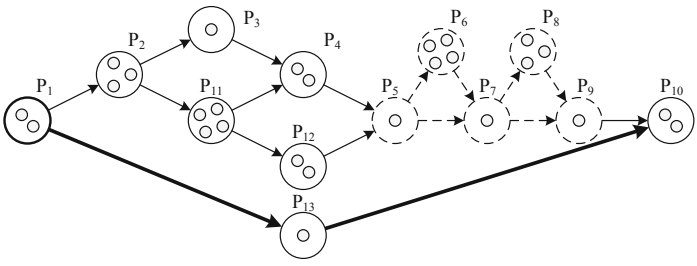

Fig. 3. A more complex questionnaire graph (Q-graph).

Table 1. The number of remaining items rem_{min}, rem_{mean}, and rem_{max} for each page of the Q-graph in Fig. 3. The estimations are based on the *max*imum, *mean*, and *min*imum prediction strategies. The table shows the pages in a reverse topological order.

Page	rem_{min}	rem_{mean}	rem_{max}	*Page*	rem_{min}	rem_{mean}	rem_{max}
P_{10}	2	2	2	P_3	8	11.5	15
P_9	3	3	3	P_{12}	7	10.5	14
P_8	6	6	6	P_{11}	11	14.5	18
P_7	4	5.5	7	P_2	11	16	21
P_6	8	9.5	11	P_{13}	3	3	3
P_5	5	8.5	12	P_1	5	11.5	23
P_4	7	10.5	14				

is by visiting the pages in reverse topological order beginning from the ending page P_{10} [19]. In a reverse topological order, each page is processed if all of its direct successor pages are processed.

For the ending page, the number of remaining items $rem(P_{10})$ is 2. This belongs to situation 1) mentioned earlier. The number of remaining items of P_9 is 3. It is computed by the number of remaining items of its direct successor page (P_{10}, 2) added with the number of its own items, 1. This belongs to situation 2). Similar holds for the number of remaining items at P_8 with $rem(P_8) = 6$. For page P_7, however, there are two direct successor pages and, therefore, two possible numbers of remaining items 6 and 3 after it. This belongs to situation 3) and the selection operator (i.e., prediction strategy) combines both values. For example, if the selection operator is the minimum function, then $rem(P_7) = min(3, 6)+1 = 4$. Taking the maximum function, the number of remaining items $rem(P_7)$ is $max(3, 6) + 1 = 7$. Table 1 shows the number of remaining items for the Q-graph for three strategies *min*imum, *mean*, and *max*imum.

Table 1 shows obvious differences especially for the pages at the begin of the Q-graph (e.g., the remaining items on P_1 and P_2). Assume a participant takes the path $(P_1, P_2, P_{11}, P_4, P_5, P_7, P_8, P_9, P_{10})$ through the survey. Depending on the differences in the numbers of remaining items, the displayed progress differs between different strategies. For example, for the minimum strategy the progress after finishing the starting page is:

$$\rho(P_1) = \rho(P_{prev}) + |P_1| \frac{1 - \rho(P_{prev})}{rem_{min}(P_1)} = 0 + 2 * 1/5 = 40\%$$

Table 2. The predicted (displayed) progress ρ_{min}, ρ_{mean}, and ρ_{max} for the different strategies *min*imum, *mean*, and *max*imum and the true progress ρ_{true} for the path $(P_1, P_2, P_{11}, P_4, P_5, P_7, P_8, P_9, P_{10})$ of the Q-graph in Fig. 3.

Page	ρ_{min} [%]	ρ_{mean} [%]	ρ_{max} [%]	ρ_{true} [%]
P_1	40.0	17.4	8.7	10.5
P_2	$40.0+3\cdot60.0/11 = 56.4$	$17.4+3\cdot82.6/16 = 32.9$	$8.7+3\cdot91.3/21 = 21.7$	26.3
P_{11}	72.2	51.4	39.1	47.4
P_4	80.2	60.7	47.8	57.9
P_5	84.1	65.3	52.2	63.2
P_7	88.1	71.6	59.0	68.4
P_8	94.0	85.8	79.5	84.2
P_9	96.0	90.5	86.3	89.5
P_{10}	100.0	100.0	100.0	100.0

For the *mean* strategy, the displayed progress after the starting page is $2 * 1/11.5 \approx 17.4\%$, and for the maximum strategy it is $2 * 1/23 \approx 8.7\%$. Table 2 contains the progresses for all strategies for all pages on the mentioned path. It further contains the *true* progress that can be easily computed since the exact number of remaining items on this path is known. The reader may identify the

(sometimes large) discrepancies between the progresses, especially at the begin of the survey. Where the progress of the minimum strategy is really fast at 50%, the progresses of the other strategies and of the true progress grow more slowly.

3 Selecting the Best Prediction Strategy

As shown in the example of the last section, different prediction strategies usually result in different predicted progresses. To allow the selection of a strategy for a given survey, we need a measure to compare the precision of them.

In the introduction of this paper, we argued that a PI should represent the *true* progress as well as possible. However, calculating the true progress needs the exact number of remaining items—that is only known *after* the participant has finished the survey. In other words, only after a participant completes the survey on a path $W = (P_1, \ldots, P_n)$, $n \geq 1$, the calculation knows the exact number of remaining items on each page $P_1, \ldots, P_n$ and can compute the true progress ρ^*.

The predicted and true progress usually have discrepancies. Given a set $\{\sqcup_1, \sqcup_2, \ldots, \sqcup_n\}$, $n \geq 1$, of prediction strategies, the *best* strategy should minimize these discrepancies. Literature proposes many measures regarding prediction accuracy and many recommendations explain in which situations a specific measure should be applied. Hyndman and Koehler [7] consider different measures of prediction accuracy in detail and provide a good overview about them. These measures have in common that they are based on the discrepancy between the predicted and the actual measured value (in our specific case, the true progress).

Imagine some people have participated in a survey. Then, the predicted/displayed and true progress are known. We can now can bring them in relation. That means, we have a value pair $\big(\rho^*(P), \rho(P)\big)$ of the true and displayed progress for each page P on all paths participants have visited. The pair $\big(\rho^*(P), \rho(P)\big)$ can be read as "on page P the true progress was $\rho^*(P)$ but the progress $\rho(P)$ was displayed". That means, all those pairs are given as a set $\mathcal{M}$. For the comparison of different strategies $\sqcup_1, \ldots, \sqcup_n$, $n \geq 2$, there is such a set for each strategy: $\mathcal{M}_1, \ldots, \mathcal{M}_n$.

If the predicted progress differs from the true progress, it results in an error $e(P) = \rho(P) - \rho^*(P)$. Notice that ρ^* and ρ have percentage scales. As a result, the error also has a percentage scale and measures based on percentage errors are applicable. Hyndman and Koehler [7] mention four typical measures of percentage errors:

1. *Mean Absolute Error (MAE)*, $\overline{|e|}$
2. *Median Absolute Error (MdAE)*, $median(|e|)$
3. *Root Mean Square Error (RMSE)*, $\sqrt{\overline{e^2}}$
4. *Root Median Square Error (RMdSE)*, $\sqrt{median(e^2)}$.

The MAE and the RMSE are common measures to evaluate prediction accuracies. The MAE uses the absolute differences between the predicted and true progresses and as a consequence treats these errors proportionally, whereas the

RMSE squares these errors. As a result, larger are weighted more strongly. The MdAE and RMdSE are similar measures, which use the median instead of the mean and as a result are more robust against outliers. Applying these measures to the errors produces fit measures for each strategy. Since all fit measures are on the same scale, they can be compared with each other. The strategy with the lowest value is the best one of the considered strategies.

If a strategy predicts the true progress exactly, each measure produces a value 0, i.e., the error between the true and predicted progress is zero. One disadvantage of the RMSE and RMdSE is that they are infinite, undefined, or skewed when all *observed* values (i.e., the true progress) are 0 or near to 0 [7]. Since the true progress has values in the range from 0 to 100%, this disadvantage does not affect them.

In fact, the MdAE and RMdSE result in almost the same values (see the appendix). It has no benefit to consider both measures in an empirical study. Our opinion is that the RMdSE should not be used since it is more difficult to compute and to interpret.

The explained approach relies on the knowledge of the true progress and, therefore, on empirical data. Unfortunately, as with any empirical study, these data is usually not available before the survey starts. It is *not* possible to select the best strategy for a survey without additional effort on collecting empirical data. To overcome this problem, data can be generated by *pilot studies*, *simulations*, or *path-explorations* of the survey for example. *Pilot studies* refer to conducting the survey with a subset of the population to obtain data for strategy selection. In *simulations* virtual participants answer the questionnaire and result in simulated data for strategy selection. In a *path-exploration*, an algorithm computes all (or most) paths of the survey and computes sample progresses for each path. But the number of such paths in adaptive surveys may be large (or exponential). All three possibilities have in common that they should represent a "realistic" usage of the different paths. Different weights exist for the paths influencing the measure and makes the generation of data more difficult. The researcher should be aware of this.

The last section considered the example Q-graph of Fig. 3 and divergences in progresses after applying three prediction strategies *min*imum, *mean*, and *max*imum (cf. Table 2). Table 3 contains the errors e between the displayed and true progress for each page of the path $(P_1, P_2, P_{11}, P_4, P_5, P_7, P_8, P_9, P_{10})$. The minimum strategy has the highest errors followed by the maximum strategy. The four measures MAE, MdAE, RMSE, and RMdSE support this observation. The RMSE has the highest value except for the minimum strategy where the median measures are a little higher. As mentioned, the MdAE and RMdSE are equal. The reason is the odd length of the path (see the appendix).

Table 3. Errors e_{min}, e_{mean}, and e_{max} between the predicted (displayed) and the true progress for the different strategies *min*imum, *mean*, and *maximum* at the path $(P_1, P_2, P_{11}, P_4, P_5, P_7, P_8, P_9, P_{10})$ of the Q-graph in Fig. 3 and the progress values of Table 2. The table further contains the *MAE*, *MdAE*, *RMSE*, and *RMdSE* values for each strategy. The mean strategy has the best values for four measures.

Page	e_{min} [%]	e_{mean} [%]	e_{max} [%]
P_1	29.5	6.9	−1.8
P_2	30.0	6.6	−4.6
P_{11}	24.9	4.0	−8.2
P_4	22.3	2.8	−10.1
P_5	21.0	2.1	−11.0
P_7	19.7	3.2	−9.4
P_8	9.8	1.6	−4.7
P_9	6.6	1.1	−3.1
P_{10}	0.0	0.0	0.0
MAE	18.2	3.1	5.9
MdAE	21.0	2.8	4.7
RMSE	20.7	3.8	7.0
RMdSE	21.0	2.8	4.7

4 Experiments

The last section described four measures that can be applied to compare different prediction strategies. A first experiment with two real and large surveys examines which measure is most suitable for comparison. In a second experiment, a simulation study with over $10k$ surveys examines if there is a single best strategy for all surveys and, if not, which trivial strategy performs best in most cases.

4.1 Experiment with Real Surveys

Our department conducts large surveys with many variables, items, and adaptivity that result in a high number of possible paths participants can "walk". We store the paths on which the participants walked through the surveys with the survey engine *Coast* [18]. With these paths, it is possible to compute the true number of remaining items for each visited page for each participant. Furthermore, for each prediction strategy, the predicted number of remaining items can be calculated in retrospect with Eq. (1) and the algorithm of Fig. 2. As a result, data sets with true and displayed progresses for each strategy and for each survey are available. This data can be used to determine the most suitable measure and the best strategy.

In general, we want to answer the following research questions with our first experiment:

Table 4. Structure and important empirical properties of *survey A* and *survey B*. $N_{Participants}$ is the number of participants, $N_{Branches}$ describes the number of branching pages in Q-graph, $|Path|$ is the empirical length of paths, N_{Items} is the empirical number of items seen, $rem(start)$ describes the number of remaining items on the starting page (values in parentheses are adjustments explained in the text). (Table taken from the conference paper [20])

	Survey A	Survey B		Survey A	Survey B
$N_{Participants}$	1041	193	N_{Items}		
$N_{Branches}$	11	38	*min*	4	6
$\|Path\|$			*mean*	246.70	290.97
min	2	2	*max*	339	377
mean	16.34	18.49	$rem(start)$		
max	25	23	$\sqcup = min$	46 (167)	7 (258)
Var	48.56	24.63	$\sqcup = mean$	115 (241)	254 (495)
			$\sqcup = max$	345 (288)	706 (700)

R1 Which measure is most suitable for comparing different prediction strategies?
R2 Is there a single best prediction strategy?

Experimental Settings. For the first experiment, we considered two of our surveys that we call *survey A* and *survey B* since their content is irrelevant. Table 4 shows characteristics of the surveys where some of the characteristics are based on empirical data. In the table, $N_{Branches}$ describes the number of pages with branches, $|Path|$ refers to the number of pages within a path, and N_{Items} is the number of items a participant has seen. Both surveys have similar structures except for $N_{Participants}$ and $N_{Branches}$. Survey *A* has more available data sets, whereas survey *B* has much more branches.

The experiment examines the previously mentioned three prediction strategies: minimum (*min*), *mean*, and maximum (*max*). The *mean* strategy does *not* represent the empirical average of items on the paths, but the selection operator used in the general algorithm.

For all three strategies and both surveys, Eq. (1) and the algorithm of Fig. 2 produced data sets. The expected remaining items on the starting page vary for both surveys (cf. Table 4, $rem(start)$) and are higher for survey *B* except for the *min* approach, which is very small with a value of 7. The values in parentheses represent adjustments on the surveys explained in the following.

Lessons Learned. During the performance of the experiments, we observed two critical characteristics of surveys resulting in bad predictions. We called them *screening paths* and *adaptive page chains*.

Screening paths are paths at the beginning of surveys in which participants receive a few key questions to determine if they are part of the survey-specific target population. Depending on their answers, the survey either continues or ends quickly. As a result, screening paths end with shortcuts (exit paths) to the ending page without many items. For example, Fig. 3 has a screening path containing only page P_1 and resulting in the exit path (P_{13}, P_{10}).

The inclusion of screening paths in progress calculation usually produces bad predictions, this was the first lesson we learned. Especially by taking the *min* strategy, the exit path has the fewest remaining items and, therefore, decreases the number of remaining items on all paths at the beginning of the survey (cf. $rem(start)$ in Table 4 or the remaining items of P_1 in Table 1 of our example survey of Fig. 3). This leads to progresses near 100% for survey B after passing the last page of the screening path. We observed no great impact of screening paths for strategies *mean* and *max*.

Adaptive page chains are subgraphs with many adaptive pages, however, each participant only sees a small number of them. The survey of Fig. 3 has a little page chain between pages P_5 and P_9 for example. Survey B consists of a lot of such pages that describe special topics. In general, each participant has only seen one or two of these approx. 30 pages. Adaptive page chains disappear for the *min* strategy in progress calculation skewing the results as most participants nevertheless see at least one page. The *max* strategy includes each adaptive page in the chain resulting in high numbers of remaining items. The *mean* strategy smooths high numbers of remaining items, however, usually only by half. In Table 4, $N_{Branches}$ indicates adaptive page chains in survey B with a value of 38 instead of 11 in survey A. It was the second lesson we learned that such chains of adaptive pages also produce bad predictions.

As a consequence, we revised our experiments with surveys A and B by eliminating screening paths from progress calculation. It is useless, otherwise, to compare the results of the different strategies. Our revisions result in new numbers of remaining items for each strategy (see $rem(start)$ in parentheses in Table 4). We left adaptive page chains as they contain important items.

Experimental Results and Discussion. Figure 4 shows the results of the first experiment. The x axis describes the true and the y axis the displayed (predicted) progress. In other words, the figure visualizes the measurement points $\mathcal{M}$ (cf. Sect. 3). The black line illustrates the true progress and a perfect prediction strategy, respectively. The *min* approach obviously results in overestimations of the progress, whereas the *max* approach results in underestimations. For survey A, *mean* has values above and below the true progress line. For survey B, the values of *mean* are all below the line. This is caused by the adaptive page chains described earlier.

Besides the measurements of true and predicted progresses, Fig. 4 contains values for measures MAE, MdAE, RMSE, and RMdSE. Actually, the MdAE and RMdSE result in equivalent values as mentioned.

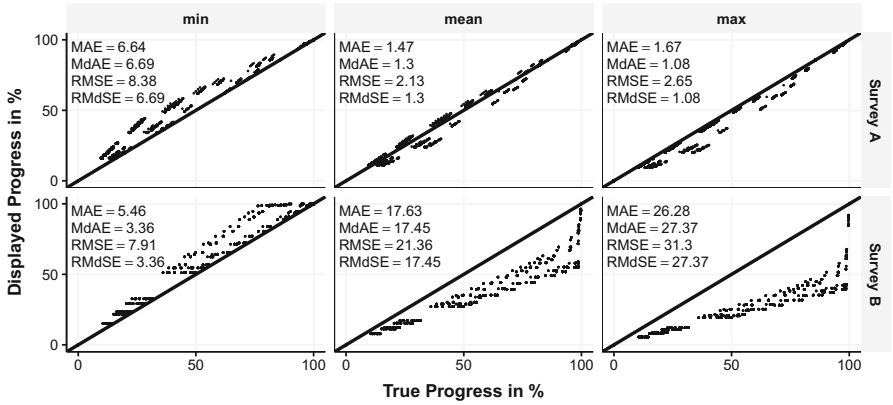

Fig. 4. Charts with displayed progresses and the computed measures MAE, MdAE, RMSE, and RMSE for the three prediction strategies *min*, *mean*, and *max* for two surveys *A* and *B*. For survey *A*, the *mean* strategy has the lowest MAE and RMSE and the *max* strategy has the lowest MdAE and RMdSE. For survey *B*, the *min* approach is the best and the *max* strategy is the worst one for all four measures. But all strategies perform worse in survey *B* (taken from the conference paper [20]).

For survey *A*, the *mean* strategy has the lowest MAE of 1.47 and RMSE of 2.13, but the *max* strategy has the lowest MdAE and RMdSE with both 1.08. The *mean* and *max* strategies seem to estimate the true progress best. The *min* strategy is the worst approach. The distribution of the points supports the result.

For survey *B*, the strategies perform contrary: the *min* approach is the best and the *max* strategy is the worst one for all four measures. All strategies perform worse for survey *B*. No strategy predicts the true progress well. Even though *min* has the lowest measures, a visual inspection of the predicted values in Fig. 4 shows that for many participants the displayed progress is near 100% even though they still have around 25% of the survey to go. This is a result of adaptive page chains, because the *min* strategy predicts that all adaptive pages will be skipped, where in most cases a participant visited at least one adaptive page. In comparison to survey *A*, all fit measures are higher, showing that progress predictions are generally worse in survey *B*. As a whole, the results show that there is no single best strategy for both surveys. We take these first results to tendentially answer research question *R2* "Is there a single best prediction strategy?" with "*No*", rather it is important to look at the characteristics of a given survey, as was seen with the adaptive page chains in survey *B*. It could also be possible that in the future a more elaborate selection strategy could be offer the best prediction for different kinds of surveys.

In our application context, high errors should be penalized more than smaller errors since higher errors have a stronger impact on the overall progress calculation and can lead to noticeable deviations from the true progress. Small errors on the contrary should be almost invisible to the participant. The RMSE is,

therefore, a good choice, because it gives large errors more weight by squaring the error. Like Fig. 4 shows, the RMSE is always the highest. The squaring of the error in RMdSE has no great effect on the resulting value. Actually, it is always close to the MdAE as mentioned before.

For survey *A*, the *mean* (MAE and RMSE) and the *max* strategies (MdAE and RMdSE) have low values. We can see in Fig. 4 that the *max* strategy has more outliers for survey *A* than the *mean* strategy. Following the above argumentation, the *mean* strategy should be used since the outliers may lead to noticeable deviations. This is supported by a higher RMSE.

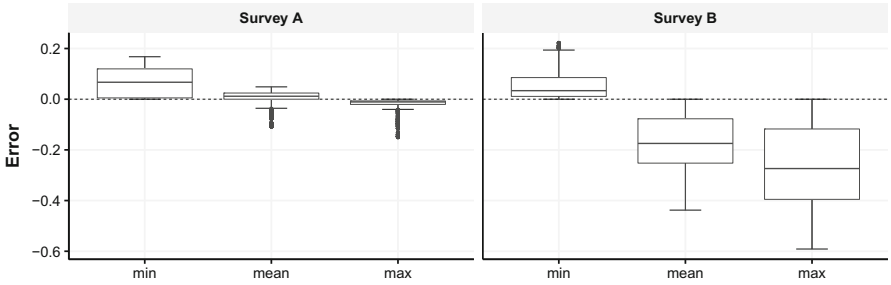

Fig. 5. Error distribution of the strategies *min*, *mean*, and *max* for surveys *A* and *B*. For survey *A*, the *mean* and *max* strategies result in errors near zero with less variance. The *min* strategy has a higher variance. In contrast, for survey *B*, the *max* and *mean* strategies have a very high variance and the *min* approach has a small variance (taken from the conference paper [20]).

Figure 5 illustrates the error distribution for all strategies in both surveys and, therefore, also the number of outliers. For survey *A*, the *mean* and *max* strategies have a median error near zero with less variance and only a minor number of outliers. Instead, the *min* strategy results in a higher variance of the error. The *max* and *mean* strategies have a very high variance with many outliers for survey *B*. The *min* approach has better results with a smaller variance. Altogether, Fig. 5 supports the previous observations and the usage of the RMSE.

Altogether, we recommend to use the RMSE for comparing different prediction strategies for PIs. It is most sensitive to high deviations. This answers research question *R1* "Which measure is most suitable for comparing different prediction strategies?".

4.2 Experiment with Simulated Surveys

The second experiment examines simulated surveys. Since our department has indeed large surveys, but does not conduct many different variants, we generated over $10k$ random surveys. It is important to note that only Q-graphs and branch

conditions with their variables were randomized and *not* complete surveys with all of their items.

The second experiment should answer research question *R2* "Is there a single best strategy?" in more detail. Including this research question, there are two questions to examine in this experiment:

R2 Is there a single best strategy?
R3 Which of the three strategies *min*, *mean*, and *max* is the best?

Experimental Settings. The experiment contains exactly 11 200 randomized surveys that are separated in 10 groups with differing numbers of pages. The page numbers N_{Pages} differ between 10 and 100 in step-size of 10. This allows a wide variety of surveys. With an increasing number of pages, we also increased the number of generated survey per group to increase the variety of different surveys. Further, the number of pages with branches $N_{Branching\,pages}$ and the number of items N_{Items} vary for each survey. The variables, the values that can be assigned to the variables, the number of clauses, and the number of axioms used in conditions are based on empirical values of surveys on our department. Our real surveys have 6 variables in conditions $N_{Variables}$ on average with 3 possible values that are assigned $M_{Possible\,values}$. The conditions are in disjunctive normal form (DNF). The average number of clauses $M_{Clauses}$ is 1.84 and the average number of axioms M_{Axioms} per clause is 1.4. Table 5 gives an overview about the surveys and their empirical properties.

Table 5. Properties of the computer-generated surveys. N_{Pages} describes the number of pages, N the number of surveys, $N_{Branching\,pages}$ is the number of pages with at least two succeeding pages, N_{Items} is the total number of items, N_{Paths} is the number of paths available, $N_{Variables}$ is the number of variables used in the conditions, $M_{Possible\,values}$ is the average of possible characteristics per variable, $M_{Clauses}$ is the average number of clauses per condition, and M_{Axioms} is the average number of axioms per clause.

N_{Pages}	N	$N_{Branching\,pages}$			N_{Items}	N_{Paths}			$N_{Variables}$			$M_{Possible\,values}$	$M_{Clauses}$	M_{Axioms}
		min	*M*	*max*		*min*	*M*	*max*	*min*	*M*	*max*			
10	400	1	2	4	200 – 1000	1	3	7	1	4	10	2.97	1.83	1.42
20	800	1	3	7	200 – 1000	1	5	24	1	6	12	2.98	1.86	1.4
30	1000	1	4	8	200 – 1000	1	5	30	1	6	12	3.01	1.85	1.44
40	1000	1	5	11	200 – 1000	1	10	176	1	6	15	3	1.86	1.42
50	1000	2	7	14	200 – 1000	1	17	234	1	7	16	3	1.85	1.42
60	1000	4	9	15	200 – 1000	1	30	339	1	8	16	3	1.84	1.42
70	1200	4	10	17	200 – 1000	1	41	452	1	8	16	3	1.85	1.42
80	1400	4	11	19	200 – 1000	1	57	571	1	8	17	3	1.85	1.41
90	1600	4	12	21	200 – 1000	1	68	712	1	8	16	2.99	1.84	1.42
100	1800	4	13	24	200 – 1000	1	82	799	1	8	18	3	1.84	1.42

It is not trivial to randomly generate realistic-looking surveys (Q-graphs). Their construction is based on the generation of so-called *Program Structure*

*Tree*s (PST) [8,9]. A PST represents a computer program as a tree with *sequence*s of *instruction*s and *bond*s (branches and loops). For its generation, the algorithm divides the given number of pages at first into a given number of branching pages (bonds) and normal pages (instructions). The second step assigns a random number of pages (instructions and bonds) to the bonds. The constructed PST is then transformed into a Q-graph by adding additional edges and joining pages. Such PSTs always result in well-structured Q-graphs. To get also some *irreducible* Q-graphs, a random number of additional edges were inserted into the Q-graph. The number of items for each page were assigned on that Q-graph randomly. Each page gets at maximum two times the average number of items per page to avoid imbalanced surveys.

Conditions on branches cause individual paths for participants. Without conditions, either all participants would take the same path or each branch is visited equal-distributed. Our random Q-graphs get conditions for each branch where each branching page has at least one default edge that is followed if none of the other branching conditions hold. The conditions are based on DNF, i.e., they consists of axioms and clauses. Each axiom consists of a variable, a comparison operator, and one possible value of the variable. The variables with their possible values are also randomly generated for each Q-graph. Sometimes, conditions are generated which never evaluate to true. But their prevention corresponds with the *SAT* problem[1] that is NP complete. We accepted this inaccuracy as improperly designed surveys.

Table 6. Descriptive statistics of the results of the second experiment. It compares the *Strategies* regarding the $RMSE$. The *mean* strategy has the lowest $RMSE$ in minimum, mean, and maximum.

Strategy	$RMSE$		
	min	M	max
min	0	10.35	53.87
mean	0	5.83	46.89
max	0	8.81	51.86

We simulated 1 000 participants for each survey. The participants were simulated by pre-assigning their answers to all variables used in the conditions in an uniform distribution. Differences in assigned answers may result in differences in paths in the surveys. All simulated participants answer the survey and use their own paths. Based on these paths, the predicted and actual progress as well as the RMSE for the three strategies *min*imum, *mean*, and *max*imum were computed.

Experimental Results and Discussion. Table 6 summarizes the results of the second experiment. For each strategy exists at least one survey for which the

[1] The *Boolean satisfiability problem* is a decision problem whether variables of Boolean formula can be replaced by *true* or *false* so that the formula evaluates to *true*.

strategy produces a perfect estimation of the true progress (the *minimum* of the $RMSE$ is 0). But for each strategy exists also at least one survey where the estimation is bad (the *maximum* of the $RMSE$ is greater than 20). The *mean* strategy seems to perform best with an average RMSE of 5.83. This is followed by the *max* (8.81) and *min* (10.35) strategies.

The averages of the three strategies show a tendency that the *mean* strategy estimates the true progress best. But sometimes outliers are hidden by the average. For this reason, we ranked each strategy for each survey from best (1) to worst (3). Figure 6 shows the three strategies and their placements, totally and relatively. The relative numbers belong to the total number of surveys.

The *mean* strategy has most placements on the first rank. It leads to the best progress estimations for about 71.9% of the surveys. If it does not perform best, it always ranked second and never third. For 28.1% of the surveys, either the *min* or *max* strategies have better RMSE. Although the *max* strategy has a better average than the *min* strategy, the *min* strategy has more rank one places than *max*. But the *min* strategy has the worst predictions in approx. 60% of the surveys. Altogether, Fig. 6 supports the conclusion drawn from comparing the average RMSE values.

The descriptive statistics show that the *mean* strategy is usually the best one of the three strategies. This answers research question *R3* "Which of the three strategies *min*, *mean*, and *max* is the best?". But the results also accentuate that non of the three strategies is best for every survey (research question *R2*). Research in future should examine strategies performing better than the strategies examined here. However, to find a strategy that performs well for each

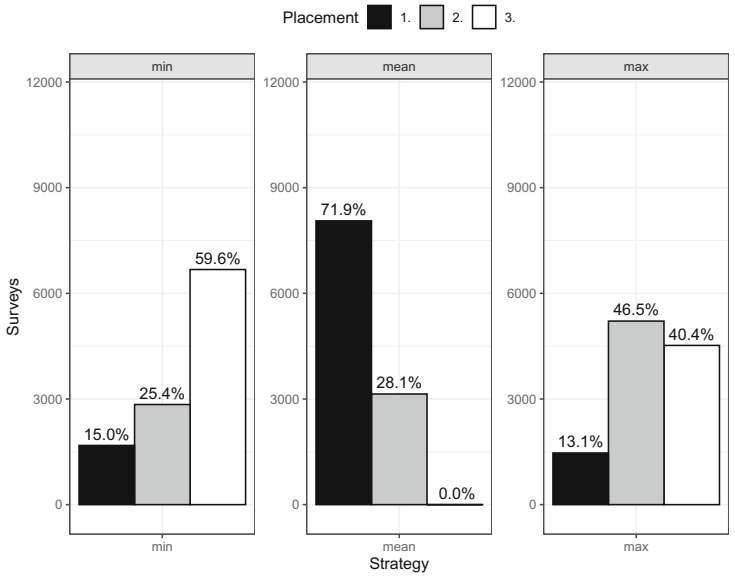

Fig. 6. Comparison of the three prediction strategies *min*, *mean*, and *max* related to their placement for a single survey. The relative values belong to the number of surveys.

survey, it becomes necessary to further investigate the factors, that influence the prediction of the progress for a given survey. We examined the describing factors of the surveys in Table 5. However, we find no significant correlations between those factors and the RMSE. These factors do not seem to describe the structure of surveys accurately in terms of progress prediction. Therefore, we need further factors that describe the structure better. One possibility is to use measures describing graph similarity. But there are a lot of such measures with individual benefits and disadvantages [11]. These measures, however, are out of the scope of this paper.

5 Conclusion

In this paper, we argued for a measure to compare prediction strategies for calculating the progress in adaptive surveys. Measures from statistics as well as experiments recommend the RMSE measure as good choice of comparing different strategies for the same survey. That strategy with the lowest RMSE is the best one. Experiments in this paper compared three standard prediction strategies: taking the minimum, average, and maximum number of items by the end of the survey. The comparison showed that the mean strategy is usually a good choice when little is known about the characteristics of the survey. In over 70% of over $10k$ simulated surveys, this strategy performs best. In all other cases, it comes in second place. However, there are surveys where even the mean strategy has poor predictions and therefore scores poorly. Our experiments with real-world and a large number of simulated surveys accentuates that *no single strategy is the best for all surveys*. The strategy is survey-dependent. Further research is necessary to provide a guide or tool for selecting the best strategy.

This study showed that using the RMSE for comparison is promising. But there is still need for empirical data. Additional research has to find ways to generate this data precisely, e.g., by further simulation or path-weighting. Furthermore, future research has to find survey characteristics that influence the accuracy of different prediction strategies. This characteristics may help to select a well-fitting strategy without having empirical data.

Sometimes it may be possible that a prediction strategy has the best RMSE within a set of strategies, but another strategy may be better in practice; for example, if simulated data does not reflect the actual population of the survey, the survey population varies over time, or there is simply an yet unknown, better strategy. We examined three basic strategies, future research should focus on finding better strategies, especially ones that can handle adaptive page chains.

Appendix

Section 3 mentioned that the MdAE and RMdSE mostly result in almost the same values. In this appendix, we are going to investigate this fact.

Assume a vector of (real) values $v = \{v_1, \ldots, v_n\}$, $n \geq 1$. The first step of computing the median $\tilde{v}$ is to order the values of v by size. By taking the

MdAE, the ordering is performed on $|v|$. For RMdSE, the ordering is performed on v^2. Since for each two arbitrary (real) values v_1 and v_2 it holds true that $|v_1| \leq |v_2| \iff v_1^2 \leq v_2^2$, it is easy to confirm that the ordering of $|v|$ is equal to the ordering of v^2. Let $(a_1, \ldots, a_n)$ be the resulting order of the absolute and $(s_1, \ldots, s_n)$ be the ordering for the squared values. The following is valid:

$$\forall 1 \leq i \leq n \colon a_i = \sqrt{s_i} \tag{2}$$

The second step of computing the median depends on the number of dimensions n of v. There are two cases:

1. n is odd. The median is the value on position $n/2$ of the ordering. It is $a_{n/2}$ for MdAE and $s_{n/2}$ for RMdSE. For RMdSE, we have to take the root median, i.e., RMdSE is $\sqrt{s_{n/2}}$. With Eq. 2 in mind, the MdAE and RMdSE are equal.
2. n is even. The median is half the sum of the values on positions $n/2$ and $n/2+1$. It is $1/2(a_{n/2} + a_{n/2+1})$ for MdAE and $1/2(s_{n/2} + s_{n/2+1})$ for RMdSE. For RMdSE, we take again the rooted median, $\sqrt{1/2(s_{n/2} + s_{n/2+1})}$. Actually, the MdAE and RMdSE are *unequal*. But on closer inspection, the value of $1/2(s_{n/2} + s_{n/2+1})$ is always between $s_{n/2}$ and $s_{n/2+1}$, and therefore, the RMdSE is always between $a_{n/2}$ and $a_{n/2+1}$ with regard on Eq. 2.

Both cases result in the following facts:

1. MdAE and RMdSE are equal if n is odd.
2. MdAE and RMdSE are almost equal if n is even and the values on positions $n/2$ and $n/2+1$ are close to each other.

References

1. Conrad, F.G., Couper, M.P., Tourangeau, R.: Effectiveness of progress indicators in web surveys: it's what's up front that counts. In: Banks, R. (ed.) Survey and statistical computing IV. The impact of technology on the survey process, vol. V, pp. 1–10. Association for Survey Computing, London (2003)
2. Conrad, F.G., Couper, M.P., Tourangeau, R., Peytchev, A.: The impact of progress indicators on task completion. Interact. Comput. **5**, 417–427 (2010)
3. Crawford, S.D., Couper, M.P., Lamias, M.J.: Web surveys: perceptions of burden. Soc. Sci. Comput. Rev. **19**(2), 146–162 (2001). https://doi.org/10.1177/089443930101900202
4. Dillman, D.A., Tortora, R.D., Bowker, D.: Principles for constructing web surveys. Technical report 98–50, Social and Economic Sciences Research Center (SESRC), Washington State University, Pullman, Washington, USA (1998)
5. Healey, B., Macpherson, T., Kuijten, B.: An empirical evaluation of three web survey design principles. Market. Bull. **16**, 1–9 (2005)
6. Heerwegh, D., Loosveldt, G.: An experimental study on the effects of personalization, survey length statements, progress indicators, and survey sponsor logos in web surveys. J. Official Stat. **22**(2), 191–210 (2006)
7. Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy. Int. J. Forecasting **22**(4), 679–688 (2006). https://doi.org/10.1016/j.ijforecast.2006.03.001

8. Johnson, R., Pearson, D., Pingali, K.: Finding regions fast: single entry single exit and control regions in linear time. Technical report TR 93–1365, Cornell University, Ithaca, NY, USA, July 1993
9. Johnson, R., Pearson, D., Pingali, K.: The program structure tree: computing control regions in linear time. In: Sarkar, V., Ryder, B.G., Soffa, M.L. (eds.) Proceedings of the ACM SIGPLAN 1994 Conference on Programming Language Design and Implementation (PLDI), Orlando, Florida, USA, 20–24 June 1994, pp. 171–185. ACM (1994). http://dl.acm.org/citation.cfm?id=178243
10. Kaczmirek, L.: Human Survey-Interaction. Usability and Nonresponse in Online Surveys. No. 6 in Neue Schriften zur Online-Forschung, Herbert von Halem Verlag, Cologne, Germany, 1st edn. (2009)
11. Koutra, D., Parikh, A., Ramdas, A., Xiang, J.: Algorithms for graph similarity and subgraph matching. Technical report, Carnegie Mellon University, December 2011
12. Li, J., Nehme, R.V., Naughton, J.: GSLPI: a cost-based query progress indicator. In: 2012 IEEE 28th International Conference on Data Engineering, pp. 678–689, April 2012. https://doi.org/10.1109/ICDE.2012.74
13. Liu, M., Wronski, L.: Examining completion rates in web surveys via over 25,000 real-world surveys. Soc. Sci. Comput. Rev. **36**(1), 116–124 (2018)
14. Luo, G.: Toward a progress indicator for machine learning model building and data mining algorithm execution: a position paper. SIGKDD Explor. Newsl. **19**(2), 13–24 (2017). https://doi.org/10.1145/3166054.3166057
15. Matzat, U., Snijders, C., van der Horst, W.: Effects of different types of progress indicators on drop-out rates in web surveys. Soc. Psychol. **40**, 43–52 (2009)
16. Myers, B.A.: INCENSE: a system for displaying data structures. SIGGRAPH Comput. Graph. **17**(3), 115–125 (1983). https://doi.org/10.1145/964967.801140
17. Myers, B.A.: The importance of percent-done progress indicators for computer-human interfaces. SIGCHI Bull. **16**(4), 11–17 (1985). https://doi.org/10.1145/1165385.317459
18. Prinz, T.M., Apel, S., Bernhardt, R., Plötner, J., Vetterlein, A.: Model-centric and phase-spanning software architecture for surveys - report on the tool Coast and lessons learned. Int. J. Adv. Softw. **12**(1&2), 152–165 (2019). iSSN 1942–2628
19. Prinz, T.M., Bernhardt, R., Plötner, J., Vetterlein, A.: Progress indicators in web surveys reconsidered – a general progress algorithm. In: Kokil, U., Ota, T. (eds.) ACHI 2019: The Twelfth International Conference on Advances in Computer-Human Interactions, Proceedings, IARIA Conference, ThinkMind Digital Library Athens, Greece, 24–28 February 2019, vol. 9, pp. 101–107 (2019)
20. Prinz, T.M., Plötner, J., Vetterlein, A.: The problem of finding the best strategy for progress computation in adaptive web surveys. In: Bozzon, A., Mayo, F.D., Filipe, J. (eds.) Proceedings of the 15th International Conference on Web Information Systems and Technologies (WEBIST 2019), Vienna, Austria, pp. 307–313, September 2019
21. Villar, A., Callegaro, M., Yang, Y.: Where am I? A meta-analysis of experiments on the effects of progress indicators for web surveys. Soc. Sci. Comput. Rev. **31**(6), 744–762 (2013). https://doi.org/10.1177/0894439313497468
22. Yan, T., Conrad, F., Tourangeau, R., Couper, M.: Should I stay or should I go: the effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys. Int. J. Public Opin. Res. **23**, 131–147 (2011). https://doi.org/10.1093/ijpor/edq046