Metacognitive Judgments in Searching as Learning (SAL) Tasks

Insights on (Mis-) Calibration, Multimedia Usage, and Confidence

Johannes von Hoyer
Leibniz-Institut für Wissensmedien IWM (Knowledge Media Research Center)
Tübingen, Germany
j.hoyer@iwm-tuebingen.de

Georg Pardi
Leibniz-Institut für Wissensmedien IWM (Knowledge Media Research Center)
Tübingen, Germany
g.pardi@iwm-tuebingen.de

Yvonne Kammerer
Leibniz-Institut für Wissensmedien IWM (Knowledge Media Research Center)
Tübingen, Germany
y.kammerer@iwm-tuebingen.de

Peter Holtz
Leibniz-Institut für Wissensmedien IWM (Knowledge Media Research Center)
Tübingen, Germany
p.holtz@iwm-tuebingen.de

ABSTRACT

Metacognitive self-assessments of one's learning performance (*calibration*) are important elements of Searching as Learning (SAL) tasks. In this SAL study, $N = 115$ participants were asked to learn for up to 30 minutes about the formation of thunderstorms and lightning by using any suitable internet resources (including multimedia resources). Participants rated their performance in comparison to other participants (*placement*), estimated the percentage of correct answers (*estimation*), and indicated their confidence in the correctness of their answers (*confidence*) in a multiple-choice knowledge test that was filled in one week before (T1) and directly after (T2) the learning phase. Participants furthermore rated the 'familiarity' of terms that do or do not exist in the context of meteorology (*overclaiming*). Learners tended to underestimate their performance at T1 and there were indicators of a potential Dunning-Kruger effect. Overall, placement and estimation ratings tended to be more accurate at T2. Surprisingly, confidence ratings increased approximately equally for correct as well as incorrect answers. A propensity for overclaiming was positively correlated with most confidence measures and the amount of time learners spent on YouTube was correlated to lower confidence scores. Implications for the design of SAL tasks and SAL studies are discussed.

CCS CONCEPTS:

• **Human-centered computing** → *User models*; **Applied computing** → *Interactive learning environments*

KEYWORDS:

Searching as learning, metacognition; overplacement; overestimation; overclaiming; self-regulated learning; judgment of learning; calibration; false certainty

SALMM '19, October 21, 2019, Nice, France

ACM ISBN 978-1-4503-6919-0/19/10…$15.00
https://doi.org/10.1145/3347451.3356730

ACM Reference Format:
Johannes von Hoyer, Georg Pardi, Yvonne Kammerer, & Peter Holz. 2019. Metacognitive Judgments in Searching as Learning (SAL) Tasks: Insights on (Mis-) Calibration, Multimedia Usage, and Confidence. In 1st International Workshop on Search as Learning with Multimedia Information (SALMM '19), October 21, 2019, Nice, France. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3347451.3356730

1 INTRODUCTION

The term *Searching as Learning* (SAL) [1] is often used to describe self-regulated learning activities on the internet. Most often, SAL research aims at predicting knowledge states or knowledge gains and/or factors that facilitate or attenuate the acquisition of knowledge [2-5].

SAL activities comprise of a sequence of information processing activities. Brand-Gruwel, Wopereis, and Walraven [6] differentiate, for example, the following five steps: (a) identification and definition of an information need and selection of corresponding keywords; (b) search for information resources, for example in the form of web pages that apart from text can also contain images and/or videos (i.e., multimedia elements); (c) evaluation of the available information; (d) extraction and processing of relevant content elements of the web pages; (e) comparison, integration, and synthesis of the extracted information. These various processing steps are accompanied by *metacognitive processes* [7], that is, reflections on one's own cognitive processes. For example, the decision to terminate a self-regulated SAL process is presumably most often based primarily on the belief that a pre-defined information need has been satisfactorily met. In this case, learners reflect on their own knowledge and thus show a special form of metacognition, namely *meta-knowledge* in the form of a judgment about the extent and quality of one's own knowledge on a subject (*judgments of learning*) [8].

According to Azevedo [9], the use of computer environments for learning challenges the metacognitive abilities of learners due to the high amount of self-regulation that is needed to fulfill learning goals. However, at the same time a computer environment can also be used as a metacognitive tool to enhance learning for example by means of supporting cognitive processes. This applies to multimedia environments, which might provide

the opportunity to offload cognitive load that is induced by the learning material as proposed by cognitive load theory [10]. As an example, consider a learner's interaction with a search engine in a SAL context: Whereas evaluating knowledge gains can be a demanding metacognitive task, encountering repeating information in search result pages (SERPs) might be a helpful cue of past learning behavior therefore facilitating judgments of learning. Hence, the main goal of the present study is to analyze the accuracy of metacognitive judgments about learning in a SAL task and to identify contributing factors to possible misjudgments. We hypothesize that SAL-processes are prone to biases, leading, for example, to overconfidence, as they have been previously identified in other self-regulated learning activities outside of the internet.

2 RELATED WORK

2.1 (Mis)-Calibration

The term *calibration* is often used for the degree to which judgments of learning correspond to objective knowledge measures [11]. *Miscalibration*, occurs whenever learners over- or underestimate their knowledge. Overestimating one's own knowledge in comparison to others is called *overplacement*, whereas the term *overprecision* is often used to describe an overestimation of subjective certainty regarding the correctness of one's answers [11].

Any form of such *overconfidence* poses a danger for self-directed learning activities, since an overestimation of one's own knowledge can hinder further study of the learning material [12, 13]. There is evidence that self-regulated learners using on-screen text compared to on-paper text, perform worse due to less accurate metacognitive judgments of their performance [14, 15]. One (often misleading) metacognitive cue that learners use to assess their knowledge and their learning progress is the perceived ease of information processing (*processing fluency*; [16]). This is particularly evident in learning with multimedia resources: Whenever learning resources contain images, learners tend to consider the material to be easier than purely textual resources conveying the same information [17, 18]. Similar effects were also reported for learning resources in the form of videos. For example, Kardas and O'Brien [19] found in a series of experiments that learners who used videos as learning material overestimated their ability to correctly perform learnable actions (e.g. throw darts), whereas learners who received the same information in the form of textual material did not. Kardas and O'Brien call this phenomenon the *illusion of skill acquisition*.

2.2 The Dunning-Kruger Effect

Another very specific form of miscalibration has come to be known as the Dunning-Kruger effect [20]. Here, the focus is on self-assessments of participants who are particularly unskilled and/or unsuccessful in a given task; methodologically, most often the lowest skilled quarter of participants is compared to the rest of the participants. In as different areas as humor, grammar, and logic [20] and contexts such as wine tasting [21], political knowledge [22], workplace end user computing [23], and reasoning in general [24] it was found that those participants who performed worst tended to vastly overestimate their performance. Although there is some research on a miscalibration of this type regarding perceived web search efficacy and actual search performance [25, 26], there have not been many attempts to study the Dunning-Kruger effect in experimental SAL-scenarios.

2.3 Overclaiming

Another interesting form of overconfidence is called *overclaiming*. Originally developed for assessing self enhancement tendencies, the *over-claiming technique* (OC) assesses participants' propensity for claiming "impossible" knowledge by asking for ratings of familiarity of existing and non-existing technical terms from the respective knowledge area [27]. Since boosting self-perceived expertise has been shown to increase the tendency to claim impossible knowledge [28] we were interested if such an effect could be observed in a SAL-task where knowledge about a topic increases after a learning phase. We also were interested in a possible link between overclaiming and other measures of (over-) confidence.

2.4 The Present Study

A major goal of the present study was to analyze leaners' (mis)-calibration and overclaiming in the context of a SAL-task, specifically, when being asked to acquire conceptual knowledge (about the formation of lightning and thunderstorms) by searching the internet. Moreover, we aim at examining whether we find evidence for a Dunning-Kruger effect.

3 METHOD

3.1 Participants

Participants were recruited via a local participant recruitment portal, which comprises of students of Tübingen University. Participants were reimbursed with 16€ per person. Of the initially 130 participants, the data from 15 participants had to be excluded because of technical (e.g., difficulties with data recordings) and organizational issues (e.g., misunderstandings regarding the procedures); hence, the data from 115 participants (96 females; M_{age} = 22.86 years; SD_{age} = 2.92) will be used in the following analyses.

3.2 Questionnaires & Data Capture

A wide range of questionnaires was used to assess participants' knowledge gains, their calibration, and psychological constructs (i.e., achievement motivation [29], cognitive reflection [30] and task engagement [31]) that can potentially influence the two. Apart from that, activity data such as information on the visited websites, mouse movements, and keystrokes was captured in a similar manner as in previous studies [2, 3]. Participants' eye movements were tracked as well. However, these data are beyond the scope of the present paper. For the present paper, we will only focus on the following measures and instruments:

3.2.1 Knowledge test. We decided to use a meteorological learning topic that had been used before in several studies on learning with multimedia [e.g., 32, 33]. Specifically, learners

were asked to learn how thunderstorms and lightning form. Information pertaining to this topic can be presented well with different representation formats such as text, pictures, schematic diagrams, or videos. Furthermore, the web provides plenty of multimodal resources on this topic. The topic can be characterized as well structured, since the physical aspects of the formation of thunderclouds and lightning are well-known and can be described comprehensively. For a profound understanding of the topic a learner must acquire procedural as well as factual knowledge pertaining to subtopics such as basic meteorology and physics as well as electricity. Based on previous work [33] we developed a 10-items multiple-choice questionnaire (MCQ) to measure knowledge one week before (T1) and directly after (T2) the learning phase. All items featured one question and four answer options of which only one was correct. Participants were informed that there was exactly one correct response for each item. For each item participants also rated their subjective confidence in the correctness of their answer on a four-point Likert-scale ranging from 1=not confident at all to 4=very confident.

Example item (Question 2): "What is a cumulonimbus cloud?"

A) Vertically extended thundercloud *(correct)*

B) A puffy cloud *(incorrect)*

C) A cotton-like cloud *(incorrect)*

D) A horizontally extended cumulus cloud *(incorrect)*

In addition, knowledge was also assessed after the learning phase using a free text essay on the topic of lightning and thunderstorms formation as well as four more complex multiple-choice questions with a varying number of correct and wrong answers to measure transfer knowledge. In the present paper, we only report the result for factual knowledge as measured by the MCQ-knowledge test described above.

3.2.2 Reasons for terminating the learning phase. After the end of the learning phase participants were asked to explain the reasons for the termination of the search process in a free text answer format.

3.2.3 Knowledge estimation and placement. For knowledge estimation we asked participants after each knowledge test at T1 and T2 to indicate their *expected number of correct answers*. They also estimated their *placement* compared to other test takers using a 10-point percentile scale from 10% to 100%

For *overplacement* and *overestimation*, we calculated the difference between the estimated and the actual number of correct answers as well as between the estimated and actual placement (defined by actual percentile rank).

3.2.4 Confidence scores. We separately calculated the *mean overall confidence* (mean of all item-based confidence ratings) as well as the *mean confidence for correct* and the *mean confidence for incorrect* answers. To measure the ability to discriminate between answers in terms of confidence ratings we additionally calculated *metacognitive sensitivity* as the difference between participants' mean confidence ratings for correct answers and for incorrect answers.

3.2.4 Overclaiming. To measure a tendency for overclaiming, we presented twelve German language meteorological technical terms (e.g. Fahrenheit, Kondensation, Mammatus) and three non-existent terms of our own invention (Thermosation, Streuungswind, Warm-Kalt-Kontrastierung) at T1 and T2. Participants rated their familiarity with those terms on a scale ranging from 0 (unfamiliar) to 6 (familiar). The instruction was as follows:

> *Please indicate for the following meteorological terms how familiar you are with them. Use the scale from 0 - 6. A rating of 0 means that you are not familiar with the term at all. A rating of 6 means that you are very familiar with the term.*

We calculated the mean familiarity rating of the three non-existing foils for each participant and measurement point as a measure for overclaiming.

3.3 Procedure

Participants were first asked to answer the factual knowledge test online approximately one week before the actual study (T1), which took place in the lab (T2). Participants assessed their expected number of correct answers and their placement and responded to the overclaiming test. In the lab, participants first worked on a working memory capacity (WMC) test and provided a free written response on what they knew about the formation of lightning and thunderstorms (not reported in the present paper). Afterwards, participants learned online for on average of 25.6 minutes (SD = 6.5) about the formation of thunderstorms and lightning by searching on the internet. They were encouraged to use any suitable (multimedia) web resources of their choice. The maximum learning time was 30 minutes; participants were free to terminate the learning phase at any point before. Participant stated their reasons for terminating in a free text box. Afterwards they were again asked to assess their number of correct answers and their placement, as well as responded to the overclaiming test again. For measuring knowledge, we presented them with the same lightning knowledge test as one week earlier and they once again worked on an essay.

3.4 Ethics Approval

Ethics approval for the study was obtained from the institute's local ethics committee (LEK 2018/068).

4 RESULTS

4.1 Knowledge Gain

Participants' knowledge scores on the 10-item MCQ-knowledge test increased significantly from 5.20 (SD=1.77) at T1 to 7.41 (SD=1.60) at T2 after the learning phase (t(255.5)=-9.85, p < .001, see Figure 1). The effect size for knowledge gain was large (d=1.29).

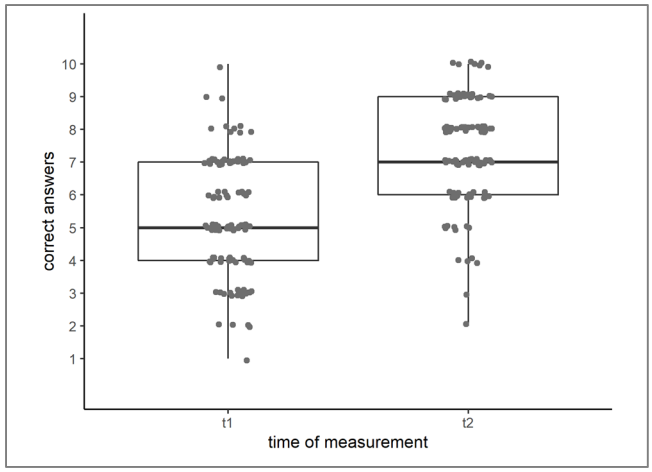

Figure 1: Mean knowledge scores for T1 and T2 (data points are scattered to prevent overplotting)

4.2 Reasons for Terminating the Learning Phase

Free text responses for the reason of termination of the learning phase were coded and aggregated. The most frequent reason for the termination of the search process was the time limit of 30 minutes (i.e., "Time ran out"). Most of the other reasons (except for "no motivation/tired/confused") are clearly related to metacognitive judgments in the form of estimates of the learners' knowledge status and evaluations of the novelty, quality, and relevance of the information that was found on the web (see table 1).

Table 1: Reasons for the termination of the SAL process derived from open answers of participants

Reasons	*Frequency*
Time ran out (30 min.)	40 (31%)
Information started to repeat itself	28 (22%)
Had found enough information	24 (18%)
Reached an understanding of the topic	21 (16%)
Information became too detailed	10 (8%)
Already used several resources	9 (7%)
No motivation/tired/confused	8 (7%)
Had found one (adequate) resource	7 (5%)
I tested myself	6 (5%)

4.3 Indicators of (Mis-) Calibration

Overall, participants appeared to be rather well calibrated in this SAL-setting as indicated by correlations between the actual and the estimated number of correctly answered questions of r=.58 (p<.001) at T1 and r=.56 (p<.001) at T2. At T1, participants displayed on average a slight propensity for underestimation (M_{diffT1}=-1.59, SD=1.64), whereas at T2 the mean difference between the actual and the estimated number of correctly answered questions was close to zero (M_{diffT2}=-0.16, SD=1.49). The difference between the two means of differences ($M_{diffT1-diffT2)}$ was significant (t(226)=-6.95, p<.001) indicating an improvement in calibration after the learning phase.

4.4 Precision of metacognitive judgments

As means of assessing the precision of metacognitive judgments of knowledge we calculated the mean overall confidence as well as the mean confidence ratings for correct and incorrect answers on an item level. Figure 2 shows the distributions of confidence ratings for correctly answered items and their change at T2. There was a significant increase in confidence with a large effect size of d=1.31 (t(978.65)=23.07, p<.001). This corresponds well to the significant average knowledge gain and the rather good calibration of participants. Surprisingly, however, a similar pattern emerged for the change in confidence for incorrectly answered questions (see Figure 3).

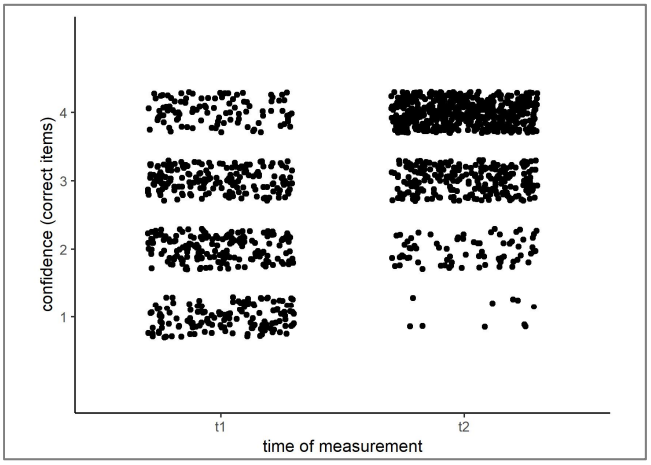

Figure 2: Confidence ratings for correctly answered items for measurement points T1 and T2 (data points are scattered to prevent overplotting)

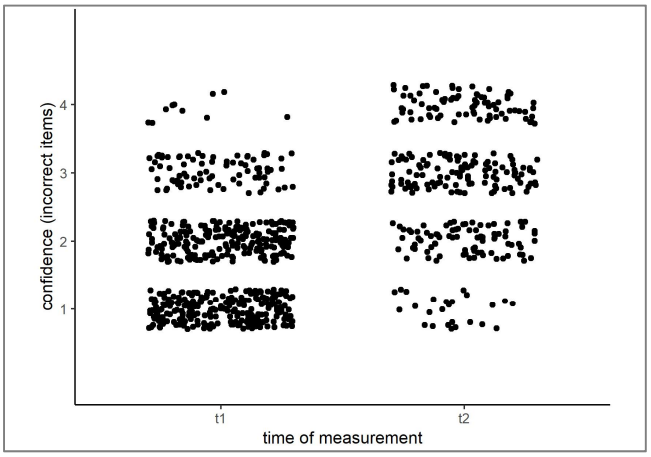

Figure 3: Confidence ratings for incorrectly answered items for measurement points T1 and T2 (data points are scattered to prevent overplotting)

Mean confidence ratings for incorrectly answered questions went up from 1.73 (*SD*=.77) to 2.82 (*SD*=.92) leading to a significant increase of *d*=1.31 (*t*(537.6)=17.46, *p*<.001).

In view of the fact that confidence scores for the incorrectly answered questions did not diminish, but increased substantially from T1 to T2, it is possible that a learning phase in a SAL-setting that is characterized by the use of multimedia resources can contribute to a *false certainty effect*. Since these rather surprising findings were exploratory, further rigorous testing of this assumption in future controlled experiments is needed to draw causal inferences. Metacognitive sensitivity did not change between T1 and T2 (t(215.68)=.03, *p*=.979; hence, participants' ability to discriminate between correct and incorrect answers as indicated by the respective confidence ratings was not affected by the learning phase.

4.5 Effects of video resources on confidence ratings

A proportion of 31% of all participants used the maximum 30 minutes of the learning phase. For the others, the average learning time was 19.89 minutes (*SD*=6.30). Only 18 participants did not use YouTube (YT) at all, while the others spent on average 39.10% (*SD*=21.43) of their learning time watching YT videos (see Figure 4).

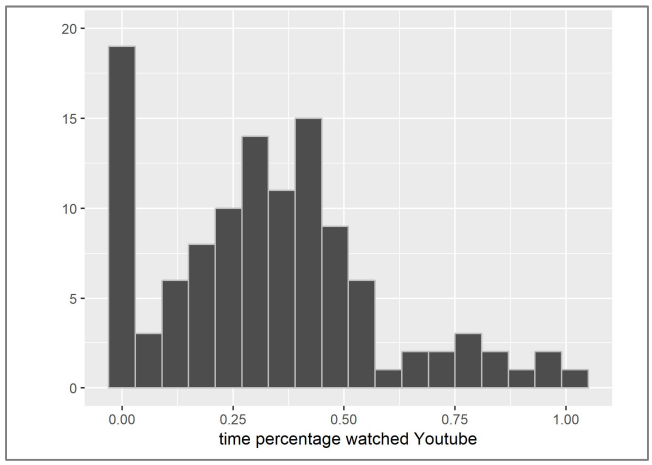

Figure 4: Distribution of percentage of time spent on YouTube during learning phase

In view of the findings on effects of video clips on confidence we mentioned earlier [19], we were interested in possible correlations between time spent watching videos and elevated confidence levels at T2. Surprisingly, we consistently found small negative correlations (Spearman's Rho) between YT usage and confidence measures and a significant positive correlation (Spearman's Rho) between YT usage and metacognitive sensitivity, albeit not all of these correlations reached statistical significance (see Table 2). There was no significant correlation between knowledge at T2 (i.e., the actual number of correct answers) and percentage of time spent on YT resources (Table 2).

Table 2: Correlations between percentage of time spent on YouTube video resource and confidence variables as well as knowledge score at T2

Variables (at T2)	*% of time spent on YouTube*	*p value*
Number of correct answers	-.03	.779
Estimated correct answers	-.21*	.043*
Estimated placement	-.16	.127
$M_{confidence}$ overall	-.20*	.047*
$M_{confidence}$ for correct items	-.14	.177
$M_{confidence}$ for incorrect items	-.28**	.007**
Metacognitive Sensitivity	.23*	.027*
Overestimation	-.13	.197
Overplacement	-.03	.759
Overclaiming	-.15	.149

Note: * indicates *p*<.05; ** indicates *p*<.01; only correlations (Spearman's Rho) between variables of matching measurement points are shown. Participants not using YT at all were excluded for this analysis resulting in *N*=97 for most variables. For variables $M_{confidence}$ *for incorrect items* and *Metacognitive Sensitivity N* dropped to 91 because of missing data.

4.6 Are there indications for a Dunning-Kruger Effect in SAL?

Although participants were rather well calibrated on measures of estimated performance, at T1, they underplaced themselves on average in comparison to other participants (M_{perc}=37%; *SD*=21%). Still, the participants in the lowest two sextiles displayed a tendency for overplacement, which is reminiscent of the original Dunning-Kruger effect [20]. However, whereas a 'typical' Dunning-Kruger effect is predominantly driven by miscalibration of the least knowledgeable participants (usually the lowest quartile), the underplacement of the best participants was at T1 even more pronounced than the overplacement of the worst performing participants (see Figure 5).

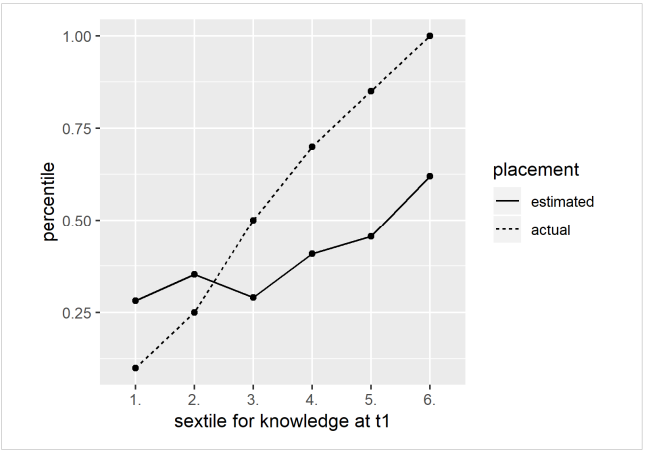

Figure 5: Placement at T1

After the learning phase (T2), participants displayed, on average, a slight tendency for overplacement (M_{perc}=62%; *SD*=19%). Participants in the lowest sextile scoring on the bottom 10% of

the knowledge test estimated their performance, on average, to be in the 48th percentile (Figure 6). Note that the amount of underplacement in the highest sextiles was comparable to the amount of overplacement in the lower sextiles. Hence, again, our findings resemble a Dunning-Kruger effect to some degree, but miscalibration is not predominantly an issue for the least knowledgeable participants.

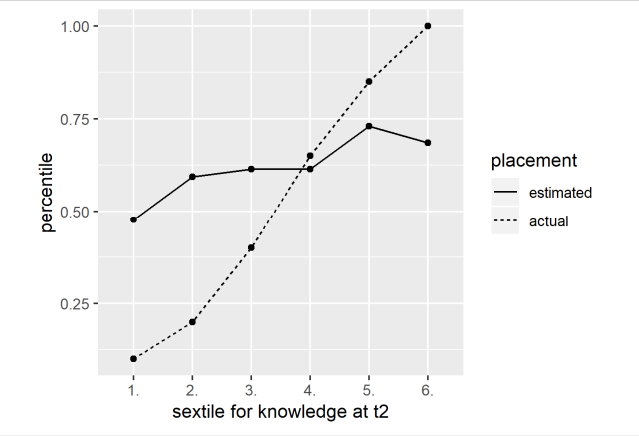

Figure 6: Placement at T2

4.7 Effects of Overclaiming

The tendency for overclaiming increased significantly from T1 (M=1.21, SD=.83) to T2 (M=1.81, SD=1.18) during the learning phase ($t(204.94) = -4.48$, $p<.05$, d=.59). This result indicates that participants claimed more impossible knowledge (higher degrees of familiarity with non-existing foil terms) after learning. Since submitting our fabricated terms to a web search resulted in no results at all, we can rule out that familiarity for those terms increased because participants were exposed to them in the learning phase. However, it must be noted that our participants had indeed been already exposed to those terms approximately one week earlier while filling out the overclaiming test at T1, which might explain the raise in familiarity.

Table 3: Correlations between overclaiming (OC), confidence variables and knowledge scores

Variables	*OC T1*	*OC T2*
No. of correct answers	.07	.04
Estimated correct answers	.25*	.15
Estimated placement	.17	.22*
$M_{confidence}$ overall	.39**	.21*
$M_{confidence}$ for correct items	.34**	.12
$M_{confidence}$ for incorrect items	.42**	.29*
Metacognitive Sensitivity	.01	-.24*
Overestimation	.20*	.11
Overplacement	.08	.09

Note: * indicates $p<.05$; ** indicates $p<.001$; only correlations between variables of matching measurement points are shown. N=114 for *Metacognitive Sensitivity* because of missing data. For all other variables N=115

Irrespective of these issues, overclaiming was at both time points associated with most measures of confidence (with the exceptions of overplacement at T1 and T2, overestimation at T2, and the average confidence for correct items at T2; see Table 3). Most interestingly, overclaiming was at both timepoints substantially correlated to the (false) confidence in incorrectly answered questions.

5 DISCUSSION & CONCLUSION

5.1 Insights for SAL Tasks

The present study provides novel insights into how metacognitive judgments influence SAL processes: Termination of a self-regulated learning activity was frequently triggered by thoughts that resembled metacognitive judgments of learning. Interestingly, one of the most frequent reasons was a perceived repetition in retrieved information. Learners in SAL-settings are free to define search engine query terms as well as to choose from the internet resources they retrieve. The feeling that the information in retrieved web resources repeats itself can result from the fact that learners have exhausted all available information on this topic on the internet, or—more likely—from the fact that learners have reached the limits of their information searching strategies and hence are unable to find additional new and relevant information.

In the present SAL-study, learners displayed a substantial knowledge gain which was accompanied by rather good calibration in terms of estimating their performance in relation to others as well as in estimating the number of correct answers. The calibration observed was comparable to previous results from studies in which students studied text material and afterwards took a knowledge test [34] as well as results from studies on long-term calibration between estimated university course performance and final exam grades [35].

Nevertheless, there was also some evidence for a Dunning-Kruger effect: At both timepoints the sextiles of participants with the lowest knowledge scores displayed a tendency for overconfidence. However, participants scoring in the highest sextiles tended to be underconfident to a similar degree. Since the common interpretation holds inefficient metacognitive calibration of the least competent participants accountable for the Dunning-Kruger effect, our findings do not quite fit into this line of explanation. As pointed out by other researchers [e.g., 36], statistical *scale end effects* are another simple and plausible explanation for the observed findings [37].

During the learning phase the participants did not only acquire knowledge; their calibration as measured by the accuracy of estimations of their test scores improved as well. Furthermore, there was, on average, an increase in the confidence ratings for correctly answered questions. Interestingly, the same effect could also be observed for the confidence in wrong answers as well. We call this effect the *false certainty effect*. Since both effects were of similar size, neither improvement nor impairment in metacognitive sensitivity (the ratio between the confidence in correct and the confidence in wrong answers) could be observed.

As was argued before [38], defining knowledge gain only by the improvement of correctly solved items of a knowledge test, lacks the important aspect of metacognitive judgments of knowledge.

Within the cue-utilization framework [39] it is assumed that metacognitive judgments of learning are based on highly salient cues. Among others, learners rely on fluency cues from processed learning materials [40]. Since learners in our SAL-setting made extensive use of multimedia resources in the form of YT videos, we would expect that those participants who spent the most time watching videos were hence exposed to many cues indicating high processing fluency, which should result in elevated confidence levels. Our results show no evidence for this assumption. Moreover, we found a trend in the opposite direction: Individuals were less confident and had better metacognitive sensitivity if they spent more time watching videos. Since confidence ratings were correlated with knowledge, one explanation for these results could be that the association between watching videos and confidence scores was being mediated by knowledge gain. However, we did neither observe a negative association between the amount of video consumption and knowledge scores at T2, nor with knowledge gain. It could be the case that many videos that were retrieved by our participants contained cues conveying impressions of poor quality; this could have led to more conservative, that is (in this case) more realistic metacognitive judgments. More research in more controlled settings is needed to further explore the relation between video and multimedia consumption during internet search and judgments of learning.

As means of exploring trait-like interindividual differences regarding confidence, we looked at correlations between participants' propensity to rate non-existing bogus terms as "familiar" (overclaiming) and confidence variables. We observed overclaiming to be consistently positively related to different measure of confidence at both points of measurement. Most notably there was a small to medium correlation with the confidence in incorrectly answered questions and a negative correlation with metacognitive sensitivity. These results indicate that interindividual differences may play a role in metacognitive judgments of knowledge.

Taken together these findings show that miscalibration can be an issue in SAL tasks, although the relations between different confidence variables appear to be complex.

5.2 Directions for Future Research

The importance of metacognitive processes for self-regulated SAL tasks cannot be overstated; yet, empirical research has only recently explicitly addressed these phenomena in detail. There are still many open questions about the intricate relationship between metacognitive judgments on the one hand and activities such as certain learning behaviors and learning outcomes on the other hand. Another important area of study for futures research are effects of multimedia usage on confidence and calibration.

Only recently SAL studies have begun to investigate the relations between learner activities and learning outcomes [1-4]. For future SAL studies, it could be worthwhile to also include indicators of metacognition as an intermediary construct between activities and outcomes, since including them allows for a more differentiated picture of knowledge. In the present study, we have integrated established instruments for assessing metacognitions into a "typical" SAL-study scenario. In future studies, think-aloud protocols [41] could be another way to further investigate participants' metacognition. In general, improving understanding in the methodology of assessing metacognitive judgments might provide useful for understanding those complex relationships.

One puzzling finding of the present study is that the confidence in wrong answers increased to approximately the same degree as did the confidence in correct answers. In view of the relationship between calibration and learning outcomes, this overconfidence in the form of false certainty poses a danger for SAL-tasks. For example, overconfidence can lead to a premature termination of the SAL process [42]. More research is needed to assess the robustness of this false certainty effect, to identify the factors that potentially contribute to it, and to find ways of preventing it in educational scenarios.

Since SAL-scenarios are increasingly integrated into formal educational contexts [43], this poses challenges for the future design of SAL-scenarios: How can we prevent overconfidence and promote improved calibration? One possible amendment could be to provide participants with automatically generated feedback on indicators of learning progress in everyday learning tasks [44, 45].

ACKNOWLEDGMENTS

The present research was funded by the German Leibniz-Association in the context of the research project "SALIENT Search as Learning – Investigating, Enhancing and Predicting Learning during Multimodal Web Search" (duration 2018-2021).

REFERENCES

[1] Pertti Vakkari. 2016. Searching as learning: A systematization based on literature. Journal of Information Science, 42(1), 7-18. http://doi.org/10.1177/0165551515615833

[2] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. 2018. (June) Predicting user knowledge gain in informational search sessions. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 75-84). ACM. http://doi.org/10.1145/3209978.3210064

[3] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. 2018. (March). Analyzing knowledge gain of users in informational search sessions on the web. In Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (pp. 2-11). ACM. http://doi.org/10.1145/3176349.3176381

[4] Nilavra Bhattacharya, and Jacek Gwizdka, 2019 (March). Measuring Learning During Search: Differences in Interactions, Eye-Gaze, and Semantic Similarity to Expert Knowledge. In Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (pp. 63-71). http://doi.org/10.1145/3295750.3298926

[5] Anett Hoppe, Peter Holtz, Yvonne Kammerer, Ran Yu, Stefan Dietze, and Ralph Ewerth. 2018. Current Challenges for Studying Search as Learning Processes. In Proceedings of Learning and Education with Web Data, Amsterdam, Netherlands, May 2018 (LILE2018), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[6] Saskia Brand-Gruwel, Iwan Wopereis, and Amber Walraven. 2009. A descriptive model of information problem solving while using internet. Computers & Education, 53(4), 1207-1217. https://doi.org/10.1016/j.compedu.2009.06.004

[7] John H. Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American psychologist, 34(10), 906-911. http://dx.doi.org/10.1037/0003-066X.34.10.906

[8] Thomas O. Nelson, and John Dunlosky. 1992. How shall we explain the delayed-judgment-of-learning effect? Psychological Science, 3(5), 317-319. https://doi.org/10.1111/j.1467-9280.1992.tb00681.x

[9] Roger Azevedo. 2005. Computer environments as metacognitive tools for enhancing learning. Educational Psychologist, 193–197. https://doi.org/10.1207/s15326985ep4004_1

[10] Richard E. Mayer. 2014. Cognitive theory of multimedia learning. In The Cambridge Handbook of Multimedia Learning, Second Edition. https://doi.org/10.1017/CBO9781139547369.005

[11] Don A. Moore, and Paul J. Healy. 2008. The trouble with overconfidence. Psychological review, 115(2), 502-517. https://doi.org/10.1037/0033-295X.115.2.502

[12] John Dunlosky, and Katherine A. Rawson. 2012. Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. Learning and Instruction, 22(4), 271-280. https://doi.org/10.1016/j.learninstruc.2011.08.003

[13] Jeri L. Little, and Mark A. McDaniel. 2015. Individual differences in category learning: Memorization versus rule abstraction. Memory & cognition, 43(2), 283-297. https://doi.org/10.3758/s13421-014-0475-1

[14] Rakefet Ackerman, and Morris Goldsmith. 2011. Metacognitive regulation of text learning. Journal of Experimental Psychology: Applied, 17, 18-32. http://dx.doi.org/10.1037/a0022086

[15] Rakefet Ackerman, and Tirza Lauterman. 2012. Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. Computers in Human Behavior, 28, 1816-1828. https://doi.org/10.1016/j.chb.2012.04.023

[16] Adam L. Alter, and Daniel M. Oppenheimer. 2009. Uniting the tribes of fluency to form a metacognitive nation. Personality and social psychology review, 13(3), 219-235. https://doi.org/10.1177/1088868309341564

[17] Allison J. Jaeger, and Jennifer Wiley. 2014. Do illustrations help or harm metacomprehension accuracy? Learning and Instruction, 34, 58-73. http://dx.doi.org/10.1016/j.learninstruc.2014.08.002

[18] Alwine Lenzner, Wolfgang Schnotz, and Andreas Müller. 2013. The role of decorative pictures in learning. Instructional Science, 41(5), 811-831. http://doi.org/10.1007/s11251-012-9256-z

[19] Michael Kardas, and Ed O'Brien. 2018. Easier seen than done: Merely watching others perform can foster an illusion of skill acquisition. Psychological science, Advance online publication. https://doi.org/10.1177/0956797617740646

[20] David Dunning. 2011. The Dunning–Kruger effect: On being ignorant of one's own ignorance. In Advances in experimental social psychology (Vol. 44, pp. 247-296). Academic Press. http://dx.doi.org/10.1016/B978-0-12-385522-0.00005-6

[21] Claudio Aqueveque. 2018. Ignorant experts and erudite novices: Exploring the Dunning-Kruger effect in wine consumers. Food Qual. Prefer. 65, (2018), 181-184. https://doi.org/10.1016/j.foodqual.2017.12.007

[22] Ian G. Anson. 2018. Partisanship, Political Knowledge, and the Dunning-Kruger Effect. Polit. Psychol. 39, 5 (2018), 1173–1192. https://doi.org/10.1111/pops.12490

[23] Shirley Gibbs, Kevin Moore, Gary Steel, and Alan McKinnon. 2017. The Dunning-Kruger Effect in a workplace computing setting. Comput. Human Behav. 72, (2017), 589–595. https://doi.org/10.1016/j.chb.2016.12.084

[24] Gordon Pennycook, Robert M. Ross, Derek J. Koehler, and Jonathan A. Fugelsang. 2017. Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. Psychon. Bull. Rev. 24, 6 (December 2017), 1774–1784. https://doi.org/10.3758/s13423-017-1242-7

[25] Barbara Coombes. 2009. Generation Y: Are they really digital natives or more like digital refugees? Synergy 7, 1 (2009), 31–40.

[26] Shinichi Monoi, Nancy O'Hanlon, and Karen R. Diaz. 2005. Online searching skills: Development of an inventory to assess self-efficacy. J. Acad. Librariansh. 31, 2 (2005), 98–105. https://doi.org/10.1016/j.acalib.2004.12.005

[27] Delroy Paulhus, and Patrick Dubois. (2014). Application of the Overclaiming Technique to Scholastic Assessment. Educational and Psychological Measurement, 74(6), 975–990. https://doi.org/10.1177/0013164414536184

[28] Stav Atir, Emily Rosenzweig, and David Dunning. 2015. When Knowledge Knows No Bounds: Self-Perceived Expertise Predicts Claims of Impossible Knowledge. Psychol. Sci. 26, 8 (2015), 1295–1303. https://doi.org/10.1177/0956797615588195

[29] Stefan Engeser. 2005. Messung des expliziten Leistungsmotivs: Kurzform der Achievement Motives Scale. Retrieved 10.08.2010 from http://www.uni-trier.de/ fileadmin/fb1/prof/PSY/PGA/bilder/Engeser__2005__Kurzform_der_AMS.pdf.

[30] Shane Frederick. 2005. Cognitive Reflection and Decision Making. J. Econ. Perspect. 19, 4 (November 2005), 25–42. https://doi.org/10.1257/089533005775196732

[31] Gerald Matthews, Sian Campbell, Shona Falconer, Lucy A. Joyner, Jane Huggins, Kirby Gilliland, . . . Joel S. Warm. 2002. Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. Emotion, 2(4), 315-340. http://dx.doi.org/10.1037/1528-3542.2.4.315

[32] Richard E. Mayer and Roxana Moreno. 1998. A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. Journal of Educational Psychology 90, 2: 312–320. https://doi.org/10.1037/0022-0663.90.2.312

[33] Florian Schmidt-Weigand and Katharina Scheiter. 2011. The role of spatial descriptions in learning from multimedia. Computers in Human Behavior 27, 1: 22–28. https://doi.org/10.1016/j.chb.2010.05.007

[34] Lin-Miao Lin, Dewayne Moore, and Karen M. Zabrucky. 2001. An assessment of students' calibration of comprehension and calibration of performance using multiple measures. Reading Psychology, 22(2), 111-128. http://dx.doi.org/10.1080/02702710130021308 3

[35] Priscilla Bell and David Volckmann. 2011. Knowledge surveys in general chemistry: Confidence, overconfidence, and performance. J. Chem. Educ. (2011). https://doi.org/10.1021/ed100328c

[36] Marian Krajc and Andreas Ortmann. 2008. Are the unskilled really that unaware? An alternative explanation. Journal of Economic Psychology, 29, 724–738. https://doi.org/10.1016/j.joep.2007.12.006

[37] Thomas Schlösser, David Dunning, Kerri L. Johnson, and Justin Kruger. 2013. How unaware are the unskilled? Empirical tests of the "signal extraction" counterexplanation for the Dunning-Kruger effect in self-evaluation of performance. J. Econ. Psychol. 39, (2013), 85–100. https://doi.org/10.1016/j.joep.2013.07.004

[38] Darwin P. Hunt. 2003. The concept of knowledge and how to measure it. Journal of Intellectual Capital, Vol. 4 No. 1, pp. 100-113. https://doi.org/10.1108/14691930310455414

[39] Asher Koriat. 1997. Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. Journal of Experimental Psychology: General, 126(4), 349-370. http://dx.doi.org/10.1037/0096-3445.126.4.349

[40] Christopher K. Hertzog, John Dunlosky, Ann E. Robinson, and Daniel Peter Kidder. 2003. Encoding fluency is a cue used for judgments about learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(1), 22-34. http://dx.doi.org/10.1037/0278-7393.29.1.22

[41] Peter Gerjets, Yvonne Kammerer, and Benita Werner. 2011. Measuring spontaneous and instructed evaluation processes during Web search: Integrating concurrent thinking-aloud protocols and eye-tracking data. Learning and Instruction, 21(2), 220-231. http://dx.doi.org/10.1016/j.learninstruc.2010.02.005

[42] Robert A. Bjork, John Dunlosky, and Nate Kornell. 2013. Self-Regulated Learning: Beliefs, Techniques, and Illusions. Annual Review of Psychology, 64,417-444. https://doi.org/10.1146/annurev-psych-113011-143823

[43] Joachim Kimmerle, Johannes Moskaliuk, Aileen Oeberst, and Ulrike Cress. 2015. Learning and collective knowledge construction with social media: A process-oriented perspective. Educational Psychologist, 50(2), 120-137. https://doi.org/10.1080/00461520.2015.1036273

[44] Mathieu d'Aquin, Alessandro Adamou, Stefan Dietze, Besnik Fetahu, Ujwal Gadiraju, Ilire Hasani-Mavriqi, Peter Holtz, Joachim Kimmerle, Dominik Kowald, Elisabeth Lex, and Susana López-Sola. 2017 (September). AFEL: Towards Measuring Online Activities Contributions to Self-directed Learning. In ARTEL@ EC-TEL. http://ceur-ws.org/Vol-1997/paper5.pdf

[45] Mathieu d'Aquin, Dominik Kowald, Angela Fessl, Elisabeth Lex, and Stefan Thalmann. 2018, (April). Afel-analytics for everyday learning. In Companion of the The Web Conference 2018 on The Web Conference 2018 (pp. 439-440). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3184558.3186206