Prospects and Pitfalls in Combining Eye-Tracking Data and Verbal Reports

Laura Helle
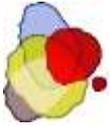
University of Turku, Finland

Abstract

It is intuitively appealing to try to combine eye-tracking data and verbal reports when investigating medical image interpretation. However, before collecting such data, important decisions must be made, including exactly when and how to collect the verbal reports. The purpose of this methodological article is to reflect on the pros and cons of different solutions and to offer some guidelines to investigators. We start by exploring the ontology of vision and speech production and the epistemology of eye movements to grasp what fixations and verbal reports actually reflect. We are also interested in the major constraints of the two systems. Second, we elaborate on two dominant investigational approaches to verbal accounts: concurrent think-aloud and Chi's explanations. Later, we move on to other approaches. Third, we present and critically evaluate studies from the literature on medical image interpretation, specifically ones that have sought to contrast or integrate eye-movement data and verbal reports. Fourth, we conclude with some practical guidelines and suggestions for further research.

Keywords: eye tracking; gaze tracking; verbal reports; think-aloud; medical images; clinical reasoning

*Corresponding author: Laura Helle, Department of Teacher Education, University of Turku, Assistentinkatu 5, 20014 Turun yliopisto, Finland. lhelle@utu.fi DOI: http://dx.doi.org/10.14786/flr.v5i3.254
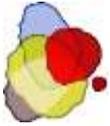
1. Introduction

The study of medical expertise in visual domains, such as radiology and dermatology, is firmly rooted in two distinct investigational approaches, both of which serve certain purposes: (a) the study of visual search or perception using eye-tracking methods (e.g., Berbaum et al., 1998; Kundel, Nodine, & Carmody, 1978; Krupinski et al., 2006; Rubin et al., 2014) and (b) the study of clinical reasoning, usually employing verbal reports (Azevedo, Faremo, & Lajoie, 2007; Lesgold, Feltovich, Glaser, & Wang, 1981; Morita et al., 2008; van der Gijp et al., 2015). Before the advent of commercial eye trackers, verbal reports were basically the only way to gain insight into diagnostic reasoning. Even today, some type of verbal report is needed because one cannot deduce from, for example, dwell times, whether a viewer actually "sees" a lesion (Berbaum, Franken, Dorfman, Caldwell, & Krupinski, 2000). A crucial part of the perceptual process is assigning meaning to what one sees (Nodine & Kundel, 1987). As for the value of eye tracking, Krupinski (2006) argued that eye tracking may be useful for developing individual eye movement profiles and for understanding the difference in performance between novices and experts. In addition, they can be useful for developing new visual search strategies.

Although studies following both lines of investigation have shown important insights, one can question whether either of the approaches alone is sufficient enough to answer important research questions. It is hard to see how medical image perception investigators are meeting the expectations of modeling, for instance, search strategies by relying on eye movement metrics alone. It also hard to see how process models can be justified based on only one source of data. As for the protocol analysts, it is odd that, for example, van der Gijp et al. (2014) conceptualized the interpretation of radiological images as a process of perception, analysis, and synthesis but methodologically relied on concurrent think-aloud techniques without the use of eye tracking.

In fact, it has been argued in the context of occupational psychology that complex cognitive work tasks should be studied by *integrating* various sources of information, including eye movement data, when appropriate, with verbal reports (Patrick & James, 2004; Gegenfurtner et al., 2017). Patrick and James (2004) stressed that process tracing involves four stages, and important decisions have to be made in each stage. The stages are the following: (1) collection of data, (2) transcription, integration, and segmentation of the data into a time-lined account, (3) coding, and (4) further analysis of the data from Stage 3 and representation of the data. In the data collection stage, one of the most critical decisions involves the timing of data collection, because verbal accounts can be collected concurrently with task performance or retrospectively. As for the transcription phase of verbal reports, the authors present the integrated actions of a person in a single table. Alternatively, one could think of either a data matrix containing a time-lined account of actions that is obtained through eye-tracking software or a set of time-stamped, transcribed videos. Stage 3 involves coding of the transcribed data either based on theoretical categories or done in a bottom-up fashion. The authors stressed that when categories are derived from a bottom-up approach, independent raters should refine categories iteratively, with some form of reliability being reported. In Stage 4, the analyst filters or expands the data using the newly acquired codes from Stage 3, and subjects the data to further analysis, whereby certain aspects of cognition are made more salient. The authors stress two points: (a) a minimum level of further analysis is whether a worker's response or solution is correct; (b) there is a need to capture and represent at a global level a person's reasoning during a scenario in relation to changes in the task and work situation.

The purpose of this methodological article is to reflect on the pros and cons of different solutions and to offer some guidelines for investigators. It is stressed that this endeavor stretches the frontiers of the field because Gegenfurtner, Siewiorek, Lehtinen, and Säljö (2013) reported in their systematic review that combining eye tracking and verbal reports remains unexplored. Also, this article is not a literature review. To review articles other than the one by Gegenfurtner et al. (2017), see Al-Moteri, Symmons, Plummer, & Cooper (2017); Blondon, Wipfli, and Lovis (2015); and van der Gijp et al. (2016). We start by exploring the ontology of vision and speech production and the epistemology of eye movements to grasp the major constraints involved in visual processing and speech production. Second, we elaborate on two dominant

investigational approaches to verbal accounts and introduce alternative approaches. Third, we present and evaluate studies from the literature on medical image interpretation that have sought to contrast or integrate eye movement data and verbal reports. Fourth, we conclude with practical solutions and some suggestions for further research.

2. Nature of the visual system: What do fixations actually reflect?

The **visual system** is the part of the nervous system that allows organisms to see. It interprets information from the environment to build a representation of the surrounding world. The visual system has the complex task of reconstructing a three-dimensional world from a two-dimensional retinal representation of that world. The performance of the visual system in a constantly-changing visual environment is remarkable. The price, however, is that approximately one-third of the cortex is needed to process visual information (Vanni, 2004).

Then, how does the visual system operate? Information from the eyes flows into the brain through the optic nerve. Information from the right visual field travels to the left optic tract. Information from the left visual field travels to the right optic tract. Each optic tract terminates in the LGN in the thalamus. The region that receives information directly from the LGN is called the V1. The Macakee Ape has over 30 cortical regions, and it is estimated that humans have approximately as many (Vanni, 2004). These areas are connected to each other by an intricate wiring containing both feedforward and feedback connections (Vanni, 2004). According to the ventral-dorsal model introduced by Goodale and Milner (1992), information flows in two directions from the primary visual cortex: (a) to the posterior parietal cortex through the dorsal stream and (b) to the inferotemporal cortex through the ventral stream. The dorsal pathway has been characterized as the action stream, a pathway concerned with converting visual inputs into motor outputs, whereas the ventral pathway provides a visual perception of objects and events in the world. Goodale (1998) stressed, however, that even a simple action such as picking up a cup a coffee requires activity in both pathways.

There are several **bottlenecks** in the visual system that stem from constraints in anatomy, attention, and working memory. First, high visual acuity is limited to the fovea, a spot on the retina. The fovea is employed for accurate vision in the direction where it is pointed. Visual acuity decreases dramatically in the parafoveal area and periphery. In eye-movement research, it is possible to capture the target of foveal inspection through fixations. Second, object recognition is limited by capacity and often attention-demanding because one cannot recognize multiple objects with more than one feature simultaneously (such as a letter T containing green and purple). Object recognition requires more than 100 ms per item, which refers to processing time instead of presentation time (Wolfe, Võ, Evans, & Greene, 2011). Third, there is a limit to focusing and shifting one's attention: people tend to move their eyes between two to four times per second when reading and conducting most visual search tasks (Salthouse & Ellis, 1980). The gaze, however, can be trained to make the best out of the few fixations: the novice's gaze is often drawn by salient, bottom-up features, whereas experts more often focus on top-down, task-relevant features, as evidenced by Bertram, Helle, Kaakinen, and Svedström (2013). (See also Wolfe, Evans, Drew, Aizenman, & Josephs, 2015). Fourth, although information flows into the system incessantly, working-memory capacity is limited to approximately four "chunks," or combinations of items, at a time (Cowan, 2010). In addition, information in one's working memory is lost quickly: according to Ericsson (2006), for tasks with response latencies of 5–10 seconds, people are able to recall their sequences of thoughts quite accurately.

For the main part, the human brain processes low-level information patterns in the environment automatically (Vanni & Heikkinen, 2015). Studies adhering to a flash-view paradigm (i.e., presenting images to participants for 20–250 ms) have shown that people can partially infer a scene without even fixating on the scene (e.g., Kirschner & Thorpe, 2006). Only a small part of the information reaching the cortex is

processed further, with storage capacity representing yet another filter. Thus, visual information processing reaching awareness is only the tip of the iceberg.

People use fixations to purposively sample information from their surroundings to reconstruct a representation of the surrounding world. However, studies adhering to a flash-view paradigm have shown that people can partially infer a scene without even fixating on the scene. Thus, there appears to be two visual pathways, coined a selective pathway involving purposive sampling and a nonselective pathway by Wolfe et al. (2011). To answer the question in the title, the sequence of fixations can be seen as reflecting a visual search through the selective pathway (i.e., attentional guidance).

3.   Speech production and verbal reports: What do verbal reports actually reflect?

3.1   Speech production

How people produce and why they produce speech is usually taken for granted. Speech has many social and cultural functions, such as signifying group identity, social grooming, settling disputes, teaching, and entertainment. Naturally, the function of each act of speech shapes **speech production**, which is a rather complex process. According to Levelt (1989, pp. 4–14), speech production involves four stages (originally conceptualized as "processing components") that depend heavily on "knowledge stores": (a) conceptualization (i.e., preverbal message generation relying on situational knowledge and content knowledge); (b) formulation, including grammatical and phonological encoding relying on lexical knowledge; (c) articulation (i.e., execution of the phonetic plan by three sets of muscles involving up to 100 different muscles) resulting in overt speech; and (d) self-monitoring (i.e., the normal components of normal language comprehension relying on lexical knowledge). Interestingly, the model includes the notion of inner speech, which is the product of the second phase. The model does not include writing as an alternation to articulation, but speech can be encoded into the visual or tactile form in addition to the auditory form. As a result, people manage to produce two to three words per second as a part of fluent conversation, and overtly naming a clear picture of an object can be initiated within 600 ms after the appearance of the picture (Levelt, Roelofs, & Meyer, 1999). In fact, the generation of inner speech may be somewhat ahead of articulation. To cope with asynchrony, it is necessary for the phonetic plan to be stored. The storage mechanism is referred to as the articulatory buffer. It is important to note that these actions tax the speaker's information-processing capacity, including working memory; in addition, speech production is delayed compared to recognition by the visual system. (Recall that object recognition requires 100 ms processing time.)

3.2   Two dominant approaches to verbal reports

In a research context, verbal reports are heavily shaped by the context in which they are produced, and verbal reports serve various functions in different research traditions. This can be highlighted by comparing two dominant approaches to verbal reports: Ericsson's protocol analysis and Chi's explanations. Ericsson and Simon's "Verbal Reports as Data" (1980), with over 13,800 Google citations and 1,619 Web of Science citations, appears to be the most influential piece of work on verbal reports. Based on Google Scholar, Ericsson has been the most active author on verbal reports over the last 30 years. Second to Ericsson and Simon's article is an article by Micheline Chi: "Quantifying Qualitative Analysis of Verbal Data: A Practical Guide." This paper has over 1,490 Google citations and over 480 Web of Science citations. These two approaches are also frequently used in the context of medical image interpretation. Therefore, Ericsson's protocol analysis and Chi's explanations deserve sections of their own.
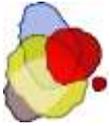
According to Ericsson (2006, p. 227), the central assumption of **protocol analysis** is that it is possible to instruct people "to verbalize their thoughts in a manner that does not alter the sequence and content of thoughts mediating the completion of a task and therefore should reflect immediately available information during thinking". Using Levelt's terminology, Ericsson is after "inner speech". In other words, the purpose is to elicit concurrent, nonreactive reports of thinking to understand expert reasoning and performance. According to the expert performance approach, the best way to obtain valid and complete traces of expert thought is to strive to produce laboratory conditions that capture "the essence of expertise," where participants perform tasks that are representative of the studied phenomenon and where verbalizations directly reflect the participants' spontaneous thoughts that are generated while completing the task. The instructions can be as follows (Ericsson & Simon, 1993, p. 376):

> "In this experiment, we are interested in what you say to yourself as you perform some tasks that we give you. In order to do this we will ask you to TALK ALOUD as you work on the problems. What I mean by talk aloud is that I want you to say out loud *everything* that you say to yourself silently. Just act as if you are alone in the room speaking to yourself. If you are silent for any length of time I will remind you to keep talking aloud."

In contrast, the goal of **Chi's explanations** (1997) is to figure out what a learner knows based on what a learner says or does and how that knowledge influences the learner's reasons. Chi avoided giving detailed instructions on how verbal reports should be elicited. Instead, she gave detailed instructions on the analysis of such reports. She stressed that one must first determine "what" the learner said (e.g., a set of propositions or concepts). However, after that, to determine the overall structure of knowledge representations, one must assess the relations between the set. For example, a learner with naïve conceptions can hold pieces of unrelated knowledge, or a learner's knowledge set can be theory-like, meaning that the reasoning can be captured by a few principles.

According to Chi (1997), the method of coding and analysing verbal data consists of the following eight steps:

1. Reducing or sampling the protocols
2. Segmenting the reduced or sampled protocols (sometimes optional)
3. Developing a coding scheme or formalism
4. Operationalizing evidence in the coded protocols
5. Depicting the mapped formalism (optional)
6. Seeking a pattern in the mapped formalism
7. Interpreting the patterns
8. Repeating the entire process, perhaps adopting a different grain size (optional).

According to Chi (1997), there are five key differences between Ericsson's protocol analysis and her verbal analysis. First, there is a clear juxtaposition in the way the verbal reports are collected. Ericsson and Simon (1993) underlined that research participants are simply verbalizing the information they attend to while generating an answer to a problem *instead of describing, explaining, justifying, or rationalizing their actions*. Second, there is a difference in focus. Ericsson and Simon (1994) were concerned with tapping the online process of problem solving or decision making, whereas Chi was interested in capturing the participants' knowledge representations. She even argued that the goal of protocol analysis is to test the a priori model rather than to uncover what the participants are actually doing. The third difference has to do with analytical procedures and workloads. According to Chi (1997), in protocol analysis, coming up with the ideal template, which requires a cognitive task analysis, represents the majority of the workload. In contrast, in verbal analysis, the referents are unknown; in self-explanation data, one must determine *what* the participant is talking about (e.g., an inference, plan, or inquiry). Fourth, the method of validation or testing is different for the two methods. In Ericsson's protocol analysis, the sequence of verbal utterances is simply compared to the ideal template. The validation of the protocol analysis is "the degree of match" between these two. In the verbal analysis method, validation is achieved by using statistical testing. For example, qualitatively different knowledge representations of different groups of participants can be checked against
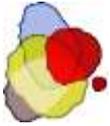
the answers to some subject-specific questions. Ericsson (2006) pointed out that task analysis *can be* applied to the analysis of think-aloud protocols. However, he added that it is also possible to examine the convergent validity established by different types of data, including reaction times, error rates, patterns of brain activation, and sequences of eye movements.

Ericsson's protocol analysis has several advantages. An obvious advantage is that Ericsson provided detailed instructions on how to collect data using the method. The other advantage is that Ericsson and Simon (1993) provided a wealth of evidence indicating that the method is not reactive (i.e., it does not alter the course of cognitive processing). The main disadvantages are the following: (a) As Morita et al. (2008) noted, medical image interpretation involves an implicit process that is difficult to verbalize; (b) thinking-aloud generally slows down performance, which may disrupt the execution of dynamic tasks in particular; (c) in certain tasks, it has been shown to alter accuracy (Russo, Johnson, & Stephens, 1989).

The disadvantage of prompting for explanations in the middle of an activity is that it has been repeatedly shown to affect behaviour in multiple ways (Ericsson & Simon, 1993). A more recent study exploring visual search behaviour on different sets of Web pages showed that prompting for explanations not only prolonged the task, but also led to more general distributed visual behaviour and the issuing of more commands to navigate within and between the Web pages. In addition, mental workload increased (Hertzum, Hansen, & Anderson, 2009). Another disadvantage of Chi's explanations is the lack of clear instructions on how to collect "explanations." If explanations are required, some form of retrospective reporting should be seriously considered.

3.3   Retrospective reporting

In fact, people can provide quite accurate retrospective reports for short tasks that take 5–10 seconds (Ericsson, 2006). The instruction can be as follows: "Can you please tell me what you were thinking during problem solving?" (van Gog, Paas, van Merriënboer, & Witte, 2005). In the context of medical image interpretation, it is also common to ask the participants to report on the findings and final diagnosis either orally or in writing. As Patrick and James (2004) pointed out, ideally, verbal reports should be collected immediately after task completion while the participant's short-term memory still holds relevant information. According to the authors, when there is a need to rely on the participant's long-term memory, some type of retrieval cues should be designed. In the case of medical images, which require more than 5–10 seconds to interpret, showing the participants a dynamic presentation of their eye movements (and keyboard movements when applicable) would seem to be a viable cuing solution.

There have been some noteworthy efforts to compare concurrent think-aloud with retrospective reports and cued reporting. In the context of troubleshooting electric circuits, van Gog, Paas, van Merriënboer and Witte (2005) conjectured that the methods would extract different types of information regarding process tracing. The authors did not report the duration of the troubleshooting tasks, but it seems safe to assume that the task durations exceeded the critical limit of 5–10 seconds. Thus, it was not surprising that the concurrent think-aloud and cued retrospective reporting involving showing the participants their eye movements and keyboard strokes resulted in more information than retrospective reporting. The remarkable finding was that the cued retrospective method resulted in less theoretical meaning-making verbalizations (why utterances), whereas the concurrent method resulted in less metacognitive utterances. Thus, in addition to the context and population, one needs to consider carefully the *type* of information one is seeking.

The advantage of retrospective reporting is that the method of verbalization does not interfere with task completion. If the task is of a very short duration (under 10 seconds), accurate reports of thinking processes can be expected. The advantage of cued retrospective reporting is that it does not interfere with task performance. However, it may be that the cues are not sufficient enough to recover all task-related information.

It is worth noting that Morita et al. (2008) showed that it can be worthwhile to triangulate verbal reports obtained through different methods. Their results indicated that experts use more conceptual words in thinking-aloud through a visual task, but they use more perceptual words when compared to novices in the writing of the report. The interpretation of the finding was that the development of expertise is based on an ability to build connections between percepts and concepts.

4. Critical examination of studies combining eye tracking and verbal reports in the context of medical image perception

Although there are many arguments for promoting the combination of eye tracking and verbal reports, combining eye tracking and verbal reports is easier said than done, as can be seen from the following studies employing concurrent think-aloud. **Concurrent think-aloud** attempts to capture nonreactive verbal reports of thinking (Ericsson, 2006). The notion of "nonreactivity" means that the execution of the primary task is not affected, except for the fact that it may be prolonged. The participants are asked to perform a task while uttering briefly what spontaneously comes to mind. In other words, it aims to "vocalize inner speech." Ericsson emphasized numerous times that participants should be talking to themselves, not explaining what they are doing or why because it has been repeatedly shown that the act of explaining can seriously interfere with the task the investigators are trying to model.

The first efforts to triangulate different sources of data obtained from different studies date back to the year 2000. Berbaum et al. (2000) were interested in conducting a congenially designed laboratory experiment to determine if satisfaction of search is because of recognition error or because of decision error by two different methods (eye tracking versus protocol analysis). The design involved inserting artificial lesions in an image to see if it decreases the detection of native lesions, indicating satisfaction of search (SOS). An earlier study employing eye tracking had indicated that inserting artificial lesions to certain images decreased the reporting of native lesions on those images. In the new experiment, Berbaum et al. (2000) discovered two important things: First, the think-aloud condition served to eliminate the satisfaction of the search effect. Second, the two methods provided contradictory results: the eye-tracking study suggested SOS was because of decision error, whereas the think-aloud study suggested that SOS was because of recognition error. The authors concluded that protocol analysis is limited in its ability to differentiate between search error and recognition error. On the other hand, there are perils in assuming that a lesion has been recognized based on dwell time alone. Thus, it was hard to reconcile the fact that the two studies produced contradictory findings. Also, the fact that the think-aloud procedure affected performance on the primary task casts doubts on the integrity of the entire study: it is difficult to argue that the think-aloud procedure was nonreactive. We speculate that the reactivity was because of the instructions given; the observers were instructed to use a finger to point to where they were looking at and to verbalize the structures and the features they were looking for. These early efforts highlight the difficulties investigators can experience in applying concurrent think-aloud and method triangulation.

In fact, the first fundamental issue to consider is whether the concurrent think-aloud condition interferes with the primary task. To our knowledge, only a single study has been conducted on this issue in the context of medical image interpretation. Littlefair, Brennan, Reed, Williams, and Pietrzyk (2012) explored whether the think-aloud condition affects pulmonary node detection; they did this using a within-subjects design with seven participants, two viewing sessions with a "wash-out period" separating them, and a set of 30 two-dimensional radiographs. Half of the radiographs contained a single artificial nodule, and the rest were non-nodular. The participants were informed that the radiographs may contain a single nodule. No time limit was set for viewing. Performance was evaluated in terms of sensitivity, specificity, and ROC measures, including multicase multireader ROC AUC analysis. In addition, the participants' eye movements were tracked to compare, for example, fixations of areas of interest and time to fixate on areas of interest.
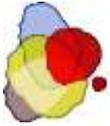
Results indicated that only half of the nodules ended up correctly localized, indicating an absence of ceiling effects. There were no differences in performance under the two conditions, with the exception of confidence ratings (in the TA condition, the subjects were less confident) and task duration. The latter result was statistically significant. The results are well-aligned with Ericsson and Simon's (1993) theoretical account.

Concurrent think-aloud has also been used, rather surprisingly, in the context of dynamic stimuli. An alternative approach situated in the context of fish locomotion can be seen in Jarodzka, Scheiter, Gerjets, and van Gog (2010). Balslev et al. (2012) used think-aloud in the context of viewing films depicting infants with seizures and conditions resembling seizures. Balslev et al. (2012) had their participants (medical students, residents, and experts) think-aloud while diagnosing the infant seizures presented in the short films, which lasted anywhere from 26–49 seconds. The films were looped and repeated until the observer wished to stop viewing. Not surprisingly, the experts scored higher in diagnostic accuracy and spent relatively more time viewing task-relevant features. A content analysis of the verbal accounts revealed that experts engaged more, in relative terms, exploring the material and spent more time building and evaluating hypotheses. This pattern, in turn, explained why the experts returned to the areas of interest. This study showed how the combined use of eye movements and verbal reports can lead to a better understanding of medical image interpretation.

Finally, Li, Pelz, Alm, and Haake (2012) attempted to integrate eye movement information completely with the concurrent verbalizations of a group of dermatologists who differed in their level of training; the groups were asked to observe 42 two-dimensional dermatological images. Subsequently, they developed a hierarchical probabilistic framework to extract unique and common eye movement patterns among multiple subjects within each expertise group. The idea was to map specific eye movement patterns to certain cognitions, such as identifying the primary morphology. Although the study is a remarkably ambitious endeavor to integrate eye tracking and concurrent verbalizations, as a process-tracing study, the work suffers from the implementation of the concurrent verbalizations. The novices were requested to provide a detailed description of the materials "as if describing to their doctors over the phone." The medical professionals were instructed to examine and describe the findings to students "as if teaching." (ibid., p. 395) These are clear violations of the principle of focusing on inner thought, and asking these questions may have seriously interfered with the primary task of image interpretation. Therefore, the mapping solution presented may have limited value.

Retrospective verbalization is a suitable option for very short tasks where there is simply not enough time to verbalize. Jaarsma, Jarodzka, Nap, van Merriënboer, and Boshuizen (2014) applied a heavily time-constrained research design. The authors presented two-dimensional microscopic images of colon tissue to a group of clinical pathologists, pathology residents, and medical students. The viewing of the images was constrained to 2 seconds. The participants' eye movements were registered, along with their post hoc verbal accounts of what they had seen. (The authors did not use the expression "retrospective think-aloud"; instead, they referred to post hoc verbalizations.) The investigators analyzed the two sources of data separately. The verbal accounts were analyzed through an elaborate content analysis. The most interesting findings related to the differences between the clinical pathologists and the residents: in their search, the clinical pathologists tended to rely on what they had already seen, further studying the image for other abnormalities, whereas the residents tended to double-check their initial findings. In their post hoc verbalizations, the clinical pathologists focused on the typicality of the tissue, whereas the residents concentrated on naming pathologies. This study showed that important insights can be gleaned by combining eye movements and a form of retrospective verbal reports.

Interestingly, not a single study could be found where the authors reported collecting the data by using **Chi's explanations** as the method. Instead, going through the literature revealed several studies where the investigators collected the data using concurrent think-aloud and then referring to Chi (1997) in the analysis phase (Azevedo et al., 2007; van der Gijp et al., 2014; van der Gijp et al., 2015). It is not unusual to find studies using concurrent verbalizations, which end up gathering explanations (e.g., Li et al., 2012; van der Gijp et al., 2015).
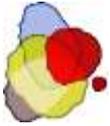
5.   Conclusions

The integration of eye tracking and verbal reports is intuitively appealing, and as illustrated in this methodological article, interesting insights can be gleaned by adopting a mixed-methods approach. However, before collecting such data, important decisions must be made. The first critical decision is timing, that is whether to collect concurrent or retrospective data. In the case of concurrent think-aloud, a decision must be made whether to follow Ericsson and Simon's or Chi's advice. In the case of retrospective reports, one must decide whether to play back the eye movements to the observer to aid the retrieval of information from long-term memory. Based on this methodological analysis, some methodological issues appear to be solved, whereas others require further investigation.

We argue that two issues are solved: First, **retrospective reporting without cuing is suitable for perceptual tasks of a very short duration** (<10 seconds). The advantages are the following: (a) one can be certain that verbalization does not interfere with the primary task; (b) the observer's verbalization is not constrained to the speed of the visual system; and (c) it is safe to assume that information is still available in short-term memory. Second, there exists a compelling body of literature indicating that for the purposes of process tracing, **Ericsson's nonreactive method is superior to the idea of soliciting direct explanations from the observers during task execution** because soliciting explanations tends to interfere with the primary task. It is hard to see what the purpose of process tracing would be if the research method results in a substantial change in the primary activity.

Other issues remain underexplored. First, when tasks are longer than 10 seconds, the pros and cons of Ericsson's concurrent think-aloud versus cued retrospective reporting need to be weighed against each other. In the study by van Gog et al. (2005) in the context of troubleshooting electric circuits, the concurrent think-aloud condition produced more theoretical expressions, which were to some extent lost in the stimulated recall condition. We emphasize that this issue has not been explored in the context of medical image interpretation. More fundamentally, there is a need for more studies to be conducted to show that perceptual tasks, such as viewing an X-ray, are not affected by the think-aloud condition. The study by Russo et al. (1989) showed that even when experimenters stick meticulously to Ericsson and Simon's instructions, the think-aloud condition may affect task performance. Russo et al. (1989, p. 758) concluded that "protocol validity should be based on an empirical check rather than theory-based assurances".

What is proposed is a research agenda with two goals: (a) to further explore if think-aloud affects performance in a range of image interpretation tasks; (b) to compare the type of information obtained by concurrent think-aloud and cued retrospective reporting with different types of material (two-dimensional, volumetric, video). It would also be useful to include observers with varying levels of experience.

Keypoints

- Before attempting to combine eye-tracking data and verbal accounts, important decisions must be made regarding the timing of the verbalizations and possible cuing.

- Ericsson's concurrent think-aloud is deemed superior to eliciting explanations from the observers during task performance.

- Retrospective think-aloud is suitable for tasks of a very short duration (<10 seconds).

- A research agenda is proposed for investigating the methodological issues that remain unsolved.

Acknowledgements

References

Al-Moteri, M. O., Symmons, M., Plummer, V., & Cooper, S. (2017). Eye tracking to investigate cue processing in medical decision-making: A scoping review. *Computers in Human Behavior*, *66*, 52-66. doi.10.1016/j.chb.2016.09.022

Azevedo, R., Faremo, S., & Lajoie, S. P. (2007, January). Expert-novice differences in mammogram interpretation. *Proceedings of the Cognitive Science Society, 29*(29).

Balslev, T., Jarodzka, H., Holmqvist, K., de Grave, W., Muijtjens, A. M., Eika, B., ... Scherpbier, A. J. (2012). Visual expertise in paediatric neurology. *European Journal of Paediatric Neurology*, *16*(2), 161-166. doi.10.1016/j.ejpn.2011.07.004

Berbaum, K. S., Franken, E. A., Dorfman, D. D., Caldwell, R. T., & Krupinski, E. A. (2000). Role of faulty decision making in the satisfaction of search effect in chest radiography. *Academic Radiology*, *7*(12), 1098-1106. doi.10.1016/S1076-6332(00)80063-X

Berbaum, K. S., Franken, E. A., Dorfman, D. D., Miller, E. M., Caldwell, R. T., Kuehn, D. M., & Berbaum, M. L. (1998). Role of faulty visual search in the satisfaction of search effect in chest radiography. *Academic Radiology*, *5*(1), 9-19. doi.10.1016/S1076-6332(98)80006-8

Bertram, R., Helle, L., Kaakinen, J. K., & Svedström, E. (2013). The effect of expertise on eye movement behaviour in medical image perception. *PloS One*, *8*(6), e66169. doi.10.1371/journal.pone.0066169

Blondon, K., Wipfli, R., & Lovis, C. (2015). Use of eye-tracking technology in clinical reasoning: A systematic review. *Studies in Health Technology and Informatics*, *210*, 90-94. doi.10.3233/978-1-61499-512-8-90

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, *6*(3), 271-315. doi.10.1207/s15327809jls0603_1

Cowan, N. (2010). The magical mystery four how is working memory capacity limited and why. *Current Directions in Psychological Science*, *19*(1), 51-57. doi.10.1177/0963721409359277

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 223-241). Cambridge, MA: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis. Verbal reports as data*. Cambridge, MA: MIT Press.

Gegenfurtner, A., Kok, E., Geel, K., Bruin, A., Jarodzka, H., Szulewski, A., & Merriënboer, J. J. (2017). The challenges of studying visual expertise in medical image diagnosis. *Medical Education*, *51*(1), 97-104. doi.10.1111/medu.13205

Gegenfurtner, A., Siewiorek, A., Lehtinen, E., & Säljö, R. (2013). Assessing the quality of expertise differences in the comprehension of medical visualizations. *Vocations and Learning*, *6*(1), 37-54. doi.10.1007/s12186-012-9088-7

Goodale, M. A. (1998). Visuomotor control: Where does vision end and action begin? *Current Biology*, *8*(14), R489-R491. doi.10.1016/S0960-9822(98)70314-8

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20-25.

Hertzum, M., Hansen, K. D., & Andersen, H. H. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, *28*(2), 165-181. doi.10.1080/01449290701773842

Jaarsma, T., Jarodzka, H., Nap, M., Merrienboer, J. J., & Boshuizen, H. (2014). Expertise under the microscope: Processing histopathological slides. *Medical Education*, *48*(3), 292-300. doi.10.1111/medu.12385

Jarodzka, H., Scheiter, K., Gerjets, P., & van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, *20*(2), 146-154. doi.10.1016/j.learninstruc.2009.02.019

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern recognition and decision making in pulmonary nodule detection. *Investigative Radiology*, *13*(3), 175-181.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*(11), 1762-1776. doi.10.1016/j.visres.2005.10.002

Krupinski, E. A., Tillack, A. A., Richter, L., Henderson, J. T., Bhattacharyya, A. K., Scott, K. M., … Weinstein, R. S. (2006). Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Human Pathology*, *37*(12), 1543-1556. doi.10.1016/j.humpath.2006.08.024

Lesgold, A. M., Feltovich, P. J., Glaser, R., & Wang, Y. (1981). *The acquisition of perceptual diagnostic skill in radiology* (No. LRDC-81/PDS-1). Pittsburgh University Learning Research and Development Center.

Levelt, W. J. M. (1989). *Speaking. From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1-38.

Li, R., Pelz, J., Shi, P., Alm, C. O., & Haake, A. R. (2012, March). Learning eye movement patterns for characterization of perceptual expertise. In *Proceedings of the symposium on eye tracking research and applications* (pp. 393-396). ACM.

Littlefair, S., Brennan, P., Reed, W., Williams, M., & Pietrzyk, M. W. (2012, February). Does the thinking aloud condition affect the search for pulmonary nodules? In *SPIE medical imaging* (pp. 83181A-83181A). Bellingham, WA: International Society for Optics and Photonics.

Morita, J., Miwa, K., Kitasaka, T., Mori, K., Suenaga, Y., Iwano, S., ... Ishigaki, T. (2008). Interactions of perceptual and conceptual processing: Expertise in medical image diagnosis. *International Journal of Human-Computer Studies*, *66*(5), 370-390. doi.10.1016/j.ijhcs.2007.11.004

Nodine, C. F., & Kundel, H. L. (1987). Using eye movements to study visual search and to improve tumor detection. *Radiographics*, *7*(6), 1241-1250.

Patrick, J., & James, N. (2004). Process tracing of complex cognitive work tasks. *Journal of Occupational and Organizational Psychology*, *77*(2), 259-280.

Rubin, G. D., Roos, J. E., Tall, M., Harrawood, B., Bag, S., Ly, D. L., ... Choudhury, R. K. (2014). Characterizing search, recognition, and decision in the detection of lung nodules on CT scans: Elucidation with eye tracking. *Radiology*, *274*(1), 276-286. doi.10.1148/radiol.14132918

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition*, *17*(6), 759-769.

Salthouse, T. A., & Ellis, C. L. (1980). Determinants of eye-fixation duration. *The American Journal of Psychology*, 207-234.

Van der Gijp, A., van der Schaaf, M. F., van der Schaaf, I. C., Huige, J. C. B. M., Ravesloot, C. J., van Schaik, J. P. J., & ten Cate, T. J. (2014). Interpretation of radiological images: Towards a framework of knowledge and skills. *Advances in Health Sciences Education*, *19*(4), 565-580. doi.10.1007/s10459-013-9488-y

Van der Gijp, A., Ravesloot, C. J., van der Schaaf, M. F., van der Schaaf, I. C., Huige, J. C., Vincken, K. L., ... van Schaik, J. P. (2015). Volumetric and two-dimensional image interpretation show different cognitive processes in learners. *Academic Radiology*, *22*(5), 632-639. doi.10.1016/j.ejrad.2014.12.015
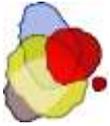
Van der Gijp, A., Ravesloot, C. J., Jarodzka, H., van der Schaaf, M. F., van der Schaaf, I. C., van Schaik, J. P. J., & ten Cate, T. J. (2016). How visual search relates to visual diagnostic performance: A narrative systematic review of eye-tracking research in radiology. *Advances in Health Sciences Education*, doi. 10.1007/s/10459-016-9698-1

Van Gog, T., Paas, F., van Merriënboer, J. J., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied*, *11*(4), 237. doi.10.1037/1076-898X.11.4.237

Vanni, S. (2004). Näkötiedon käsittely aivokuoressa [Processing of visual data in the cerebral cortex]. *Duodecim, 120*, 2653-2662.

Vanni, S., & Heikkinen, H. (2015). Onko aivoissamme käyttämätöntä kapasiteettia? [Is there unused capacity in our brain?] *Duodecim, 131*, 1644-1649.

Wolfe, J. M., Evans, K. K., Drew, T., Aizenman, A., & Josephs, E. (2015). How do radiologists use the human search engine? *Radiation Protection Dosimetry*, 501. doi.10.1093/rpd/ncv501

Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*(2), 77-84. doi.10.1016/j.tics.2010.12.001