Self-monitoring in graduate medical education: An opportunity for the think aloud interview

William Rainey Johnson, MD, MEd
LCDR MC(UMO) USN

Thesis submitted to the Faculty of the
Uniformed Services University of Health Sciences
In partial fulfillment of the requirements for the degree of
Master of Health Professions Education
2022

Disclaimer: *The opinions and assertions expressed herein are those of the author(s) and do not reflect the official policy or position of the Uniformed Services University of the Health Sciences or the Department of Defense. References to non-Federal entities or products do not constitute or imply a Department of Defense or Uniformed Services University of the Health Sciences endorsement.*

Conflicts of Interest: *Neither I nor my family members have a financial interest in any commercial product, service, or organization providing financial support for this research.*

Distribution Statement

Distribution A: Public Release.

The views presented here are those of the author and are not to be construed as official or reflecting the views of the Uniformed Services University of the Health Sciences, the Department of Defense or the U.S. Government.

April 12, 2022

APPROVAL SHEET

Title of Thesis: Self-monitoring in graduate medical education: An opportunity for the think aloud interview

Name of Candidate: W. Rainey Johnson, MD, MEd, Master of Health Professions Education

3/28/2022

THESIS AND ABSTRACT APPROVED:

DATE: 3/28/2022

[redacted] 3/28/2022

Ronald M. Cervero
CENTER FOR HEALTH PROFESSIONS EDUCATION
Committee Chairperson

[redacted] 3/28/2022

Steven J. Durning, MD, PhD
CENTER FOR HEALTH PROFESSIONS EDUCATION
Thesis Advisor

[redacted] 3/28/2022

Anthony R. Artino, Jr., PhD
School of Medicine and Health Sciences
The George Washington University
Committee Member

Acknowledgements

I would like to thank my health professions education research and mentor team, especially Steve Durning and Tony Artino, for their continual and ongoing support and guidance. Additionally, I am very thankful for the support of my various commands and, in particular, the individuals of my professional community that have gone above and beyond to encourage me in this endeavor, namely Dr. Adam Barelski, CAPT Hugh Dainer, and COL Josh Hartzell.

Dedication

My efforts are meaningless without the support of my family.

To Sarah and the rest of my loving family (Chase, Cindy, Fischer, Giulia, Philly, and Robbie) for never-ending commitment and support that make everything I do possible.

Copyright Statement

The author hereby certifies that the use of any copyrighted material in the thesis manuscript entitled:

"Self-monitoring in graduate medical education: An opportunity for the think-aloud interview"

is appropriately acknowledged and, beyond brief excerpts, is with the permission of the copyright owner.

William Rainey Johnson
Center for Health Profession Education
Uniformed Services University
Date 04/05/2022

Abstract

Accurate physician self-monitoring, in-the-moment self-awareness of performance, may be an important factor for improving patient safety, by reducing diagnostic and therapeutic reasoning errors. While physicians and physicians-in-training have repeatedly demonstrated poor accuracy of global self-assessments, which are assessments removed from the context of a specific task, regardless of any intervention. Self-monitoring offers a promising alternative to global self-assessment. In this thesis, we will first explore the current role of self-monitoring in graduate medical education, identifying gaps in current literature and opportunities for future work. Next, we will describe the think aloud protocol in a qualitative methodology review, highlighting best-practices and common pitfalls for researching self-monitoring to research the more obscure aspects of self-monitoring. Taken together, this work will contribute to the growing understanding of self-monitoring and lay the groundwork for future understanding using the think aloud protocol with long-term goal of contributing to improved patient safety.

Table of Contents

Chapter 1: Introduction

Diagnostic and therapeutic reasoning errors threaten patient safety (Bates & Singh, 2018). Understanding and improving physicians' self-monitoring, that is their in-the-moment self-awareness of performance, may be an important factor in reducing these diagnostic and therapeutic reasoning errors. In this thesis, we will explore the current role of self-monitoring in graduate medical education and detail an underutilized method for researching self-monitoring.

Resident and fellow physicians graduate their respective programs and begin making diagnostic and therapeutic decisions independently as part of their new clinical responsibilities. Diagnostic reasoning and therapeutic reasoning errors account for approximately 15% of patient safety events, affecting over 12 million people in the United States annually (Bates & Singh, 2018; Singh & Graber, 2015). Clinical providers struggle to keep up with evolving medical knowledge and skills after training, and failures to stay current can negatively impacting patient outcomes (Day et al., 1988). Moreover, clinical practice is often complex and time pressured, filled with inherent uncertainty (Pomare et al., 2019). While some interventions, namely point of care resources and computer-assisted triggers, have initially helped to reduce diagnostic and therapeutic reasoning error, these positive impacts wane overtime due to alert fatigue (Abimanyi-Ochom et al., 2019). Ultimately, however, the only person present for every diagnostic and therapeutic reasoning decision is the responsible provider. Physicians are believed to naturally engage in self-monitoring (Ilgen et al., 2020) in the workplace environment. So, if physicians could accurately self-monitor and, aligning the self-judgments of

their performance with their true competence, then the healthcare system may be able to target successful interventions to reduce diagnostic and therapeutic reasoning errors.

Self-monitoring should not be confused with self-assessment, which is ill-defined, but generally meant to represent a personal reflection or summary of one's overall performance in a particular area typically related to knowledge, skill, or attitude (Eva & Regehr, 2009). Physicians and physicians-in-training have a predilection for *self-monitoring*, rather than *self-assessment*. Physicians and physicians-in-training have repeatedly demonstrated poor accuracy of global *self-assessments*, which are removed from the context of a specific task, regardless of any intervention (Davis et al., 2006; Eva & Regehr, 2005). For example, if a trainee were asked to assess her ability to insert an arterial line in a critically ill patient, she would likely provide an inaccurate assessment of her capabilities. On the other hand, research suggests that medical students and residents can accurately *self-monitor,* where self-monitoring is defined as in-the-moment self-awareness of performance (Brydges et al., 2016; Eva & Regehr, 2009; Gingerich et al., 2011; McConnell et al., 2012; Quick et al., 2017; Trajkovski et al., 2012). Imagine the same trainee is asked to evaluate her execution of an arterial line placement immediately after inserting one in a critically ill patient. Tying the self-appraisal temporally and contextually to the action helps her to evaluate her ability more accurately. In other words, the physician in the example above would likely be able to accurately self-monitor. Unfortunately, despite this idea, there is a paucity of literature on physician self-monitoring in non-procedural activities that require both non-analytic (e.g., pattern recognition) reasoning and more deliberate, analytic reasoning.

Educational theory

This thesis is grounded in social learning theories, namely social cognitive self-regulated and situated cognition theories. Self-regulated learning theories provide frameworks for developing physicians that can actively recognize their limitations, leverage peer and clinical support tools, and sculpt their learning to meet the demands of an evolving and often uncertain clinical practice.

The social-cognitive perspective on self-regulated learning theory includes at least four core features: goal setting, motivation, self-monitoring, and cyclical feedback (Cleary et al., 2013). A recent systematic review found that goal setting and motivation commonly appeared in the undergraduate medical education and graduate medical education literature; however, the process of regulatory appraisal, which encompasses self-monitoring and cyclical feedback, rarely appeared (Houten-schat et al., 2018). Similarly, a scoping review of self-regulated learning in medical students identified only four studies that explored situated self-monitoring, where students self-monitoring was observed in the workplace environment, that is, the learners were *situated* within the environment of practice (Cho et al., 2017). Self-monitoring situated within the clinical context is the very place where physicians must practice and master the skill of accurately aligning their self-judgment of performance and true competence and recognizing their limitations.

Social cognitive views of self-regulated learning theory postulate that learning occurs through four core features that happen iteratively before, during, and/or after a specific task. These four core features are motivation, goal setting, self-monitoring, and feedback loop (Artino et al., 2015; Cleary & Sandars, 2011). While the learner plays a central role in seeking

guidance, collaborating on goal setting, encouraging feedback, and monitoring needs and progress, social and environmental interactions also influence each of these processes (Houten-schat et al., 2018; Özdemir, 2011). This thesis focuses on one of the core processes of self-regulated learning – self-monitoring.

Given the criticality of the contextual and temporal associations of action and self-monitoring, self-monitoring depends on the social and environmental interactions emphasized in situativity theory. Situativity theory suggests that knowledge forms at the intersection of the learner and environment (Durning & Artino, 2011; Young, 1993). In GME, learners must develop knowledge, skills, and attitudes at the highest level of Miller's pyramid (Kogan & Holmboe, 2017). Learning requires the correct context. A resident cannot become independently proficient in arterial line placement without placing arterial lines in real work scenarios, where each placement is partially dependent on the equipment available, the patient's position, the provider's self-efficacy, and the educational culture of the intensive care unit. Simulated scenarios are still useful for learning and the higher fidelity the simulation has, the more reliably the learning will translate into independent practice. Self-monitoring is the same. First, self-monitoring requires context. That is one of the features that differentiates self-monitoring from self-assessment (Eva & Regehr, 2009). Second, self-monitoring, as viewed through social cognitive learning theory, depends on social and environmental interactions. Therefore, self-monitoring outside real work scenarios may be different from self-monitoring in simulated scenarios. And, the higher the fidelity of the simulated scenarios, the more likely the self-monitoring will mimic self-monitoring in a real work environment. In this thesis, we will evaluate how situated learning environments are used in self-monitoring research and consider

how future research could prioritize the situated learning environments or, at the very least, substitute the best alternative – high fidelity, authentic simulation.

The purpose of this thesis is to explore the documented existence and use of self-monitoring in graduate medical education and to describe a detailed approach to the think aloud interview, a technique for understanding thinking, such as the process of self-monitoring. In addressing these two aims, we hope to lay the foundation for improved understanding and intentional development of self-monitoring in graduate medical education.

Chapter 2: A scoping review of self-monitoring in graduate medical education

Authors

W Rainey Johnson, MD, MEd, Steven J Durning, MD, PhD, Rhonda J Allard, MLIS, Adam M Barelski, MD, Anthony R Artino, Jr., PhD

Acknowledgments

Thank you to Drs. Ryan Brydges and Timothy Cleary for reviewing our list of included articles.

Abstract

Background: Physicians and physicians-in-training have repeatedly demonstrated poor accuracy of global *self-assessments*, which are assessments removed from the context of a specific task, regardless of any intervention. Self-monitoring, an in-the-moment self-awareness of one's performance, offers a promising alternative to global self-assessment. The purpose of this scoping review is to better understand the state of self-monitoring in graduate medical education (GME).

Methods: We performed a scoping review following Arksey and O'Malley's six steps: identifying a research question, identifying relevant studies, selecting included studies, charting the data, collating and summarizing the results, and consulting experts. Our search queried Ovid Medline, Web of Science, PsychINFO, Eric, and EMBASE databases from 1 Jan 1999 to 16 April 2019.

Results: The literature search yielded 4128 unique articles. The authors identified 58 articles for inclusion. More than 20 different terms were used to describe self-monitoring and only seven studies (13%) provided a definition for the equivalent term. The research focused primarily on post-performance self-judgments of a procedural skill (n=23, 40%). Regardless of task, studies

focused on self-judgment (n=47, 81%) and accuracy of self-monitoring (n=41, 71%). Most self-monitoring was conducted post-task (n=50, 86%).

Conclusion: Self-monitoring is a complex phenomenon that seems promising as a research focus to improve clinical performance of trainees in GME and beyond. The landscape of current literature on self-monitoring is sparse and heterogeneous, suffering from a lack of theoretical underpinning, inconsistent terminology, and insufficiently clear definitions.

Introduction

Physicians and physicians-in-training have repeatedly demonstrated poor accuracy of global *self-assessments*, which are removed from the context of a specific task, regardless of any intervention (Davis et al., 2006; Eva & Regehr, 2005). Self-assessment as a construct is ill-defined, generally representing a personal reflection or summary of one's overall performance in a particular area typically related to knowledge, skill, or attitude (Eva & Regehr, 2009). Yet, the idea that practicing physicians should learn to recognize their performance limitations and opportunities for growth is appealing and appears in the guidance of numerous accreditation bodies (ACGME & ABIM, 2013; ACGME & ABS, 2015; Frank et al., 2015; *Standards for Assessment and Accreditation of Primary Medical Programs by the Australian Medical Council 2012*, 2012). After all, the only person who is consistently "with" a practicing physician is the physician themself.

Self-monitoring, an in-the-moment self-awareness of one's performance, offers a promising alternative to self-assessment (Eva & Regehr, 2011; McConnell et al., 2012). For example, if a trainee were asked to assess their ability to insert an arterial line in a critically ill

patient, they would likely provide an inaccurate assessment of their capabilities. On the other hand, research suggests that undergraduate medical education trainees can accurately *self-monitor* (Eva & Regehr, 2011; McConnell et al., 2012). In other words, a trainee can accurately evaluate her performance of an arterial line placement in Ms. Jones, a 72-year-old critically ill patient, *immediately before, during, or after completing it* at 0115 during a shift in the intensive care unit.

The literature on self-monitoring for post-graduate physician trainees is also promising, albeit a little less clear (Brydges et al., 2016). Like the medical student placing the arterial line in Ms. Jones, tying the self-awareness of performance temporally and contextually to the action may help graduate medical education (GME) trainees to more accurately evaluate their abilities. In other words, much like the novice medical student, GME trainees may be able to accurately self-monitor. Unfortunately, despite the intuitive appeal of this idea, the literature on trainee self-monitoring in GME has not been rigorously appraised.

Resident and fellow physicians graduate their respective programs and assume roles in clinical care, making diagnostic and therapeutic decisions independently. However, diagnostic and therapeutic errors still account for approximately 15% of patient safety events, negatively affecting over 12 million people in the United States annually (Bates & Singh, 2018). What is more, clinicians struggle to keep up with evolving medical knowledge and skills after training, and failures to stay current can negatively impact patient outcomes (Day et al., 1988). Finally, clinical practice is often complex and messy, filled with inherent uncertainty that requires critical self-monitoring (Pomare et al., 2019). Ultimately, the only person present for every diagnostic and therapeutic decision is the responsible clinician.

Self-monitoring is considered an essential feature of many self-regulated learning (SRL) theories (Artino et al., 2015; Cleary et al., 2013). As a group, SRL theories provide a framework for understanding developing physicians who can actively recognize their limitations, leverage peer and clinical support tools, and sculpt their learning to meet the demands of an evolving and often uncertain clinical practice. The social-cognitive perspective on SRL includes at least four core features: goal setting, motivation, self-monitoring, and cyclical feedback (Cleary et al., 2013). A recent systematic review found that goal setting and motivation commonly appeared in the undergraduate medical education and GME literature; however, the process of regulatory appraisal, which encompasses self-monitoring and cyclical feedback, rarely appeared (Houten-schat et al., 2018). Similarly, a scoping review of SRL in medical students identified only four studies that explored situated self-monitoring, where students self-monitoring was observed in the workplace environment (Cho et al., 2017). Yet, self-monitoring situated within the clinical context is the very place where physicians must practice and master the skill. Consistent with social cognitive learning theories, self-monitoring encompasses self-observation (i.e., recognizing one's thoughts, attitudes, or behaviors), self-judgment (i.e., grading accuracy/appropriateness of one's own thoughts, attitudes, or behaviors), and self-reaction (i.e., reinforcing, planning, or implementing change in thoughts, attitudes, or behaviors) (**Figure 1**) (Zimmerman & Schunk, 1989). Relevant to task performance, self-monitoring can occur immediately before, during, or after. The significance of the timing of self-monitoring and the particular type of self-monitoring being used is uncertain.

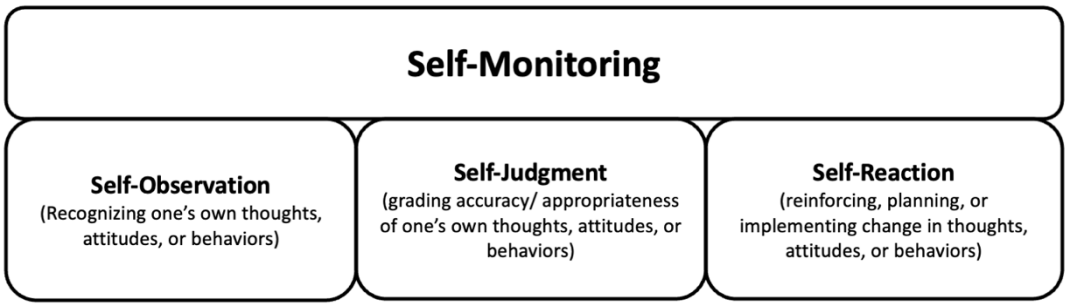

Figure 1. Shows three types of self-monitoring.

The purpose of this study is to better understand the state of self-monitoring in GME; specifically, we wanted to know how self-monitoring is incorporated in GME—how self-monitoring is encouraged, characterized, measured, and taught. For our purposes, "encouraged" means actions intended to increase the frequency of self-monitoring; "characterized" describes the process of engaging in self-monitoring; "measured" refers to the accuracy of self-monitoring; and "taught" represents efforts to improve the quality of self-monitoring. In doing so, we hope to discover best-practices within GME to promote self-monitoring, identify the gaps in the empirical literature, and recommend future directions for self-monitoring research within GME.

Methods

As a type of knowledge synthesis, scoping reviews are best for synthesizing a large volume of literature to understand its current state for the overarching purpose of identifying gaps in knowledge or practice that are ripe for future research (Thomas et al., 2017). Therefore,

we performed a scoping review to address our broad research question, following Arksey and O'Malley's six steps: identifying a research question, identifying relevant studies, selecting included studies, charting the data, collating and summarizing the results, and consulting key stakeholders or experts (Arksey & O'Malley, 2005; Levac et al., 2010; Thomas et al., 2017).

Identifying a research question: We conducted preliminary literature searches and had collaborative discussions to hone our research question: How is self-monitoring incorporated (e.g., characterized, encouraged, measured, taught) in graduate medical education?

Identifying relevant studies: We conducted a literature search in April 2019 in collaboration with a research librarian (RA). The search queried Ovid Medline, Web of Science, PsychINFO, Eric, and EMBASE databases. These specific search strategies for each database can be found in appendix 1. We intentionally created an inclusive, broad search strategy given the heterogeneity of terms used to describe self-monitoring and uncertainty of volume of literature for a GME specific population based on our preliminary search strategies.

Selecting included studies: We removed duplicates, publications before 1999, and non-primary literature, such as reviews, perspectives, and opinion pieces. Two authors (WRJ, SJD) reviewed titles, abstracts, and indexed descriptive fields to exclude results without a GME population, without content likely related to self-monitoring, and with insufficient information. Any disagreements were

included for full text review. Each full text article was reviewed by two authors (WRJ & AMB, ARA, or SJD). Any disagreements were discussed as a group with a minimum of three authors until consensus was reached. During the full text review, we looked at potentially relevant citations to ensure that they had been captured in our search results. Where a reference was missing, we evaluated the full text of the reference for potential inclusion.

Charting the data: We charted the data using a coding sheet that we developed iteratively by preliminary coding a subset of included articles and group discussions among the co-authors, grounded in a social-cognitive theory of SRL theory and, specifically, self-monitoring. We applied the final coding sheet to each included article (Appendix 2). Given the heterogeneity of types of articles, we assessed the quality of each study using two different, well-studied quality metrics – the Medical Education Research Study Quality Instrument (MERSQI) and Standards for Reporting Qualitative Research (SRQR) (Cook & Reed, 2015; O'Brien et al., 2014). Quantitative studies were evaluated with MERSQI, qualitative studies with SRQR, and mixed methods studies with both. Given that not all fields were relevant for each study, we decided to report quality scores as a percentage of the maximum possible points an article could receive based on the applicable fields. One author (WRJ) coded each article. A second author (AMB, ARA, or SJD) reviewed each article and assigned coding. Any discrepancies

were resolved through discussion with a minimum of three authors present (ARA, SJD, WRJ).

Collating and summarizing the results: One author (WRJ) collated and summarized the results to discuss with the co-authors. Together, we reviewed the findings and, through group consensus, determined which findings to include as results. Based on these results, we collaborated to determine the key take away messages from our findings and value added to the self-monitoring literature. Throughout the collation process, we attempted to practice reflexivity, recognizing our biases that may impact our judgments.

Consulting key stakeholders or experts: Throughout the review process, we consulted stakeholders, experts in the field of SRL, to discuss our research question and review our included articles to minimize the risk of relevant missing publications. We consider experts individuals that had published multiple peer-reviewed journal articles in the field of self-regulated learning in medical education. Any suggested articles were reviewed or re-reviewed for inclusion, depending on whether they had been included in our search results.

Results

The literature search yielded 4128 articles from the five different databases, after removing duplicates. Through iterative reviews of the abstracts and full-text articles, we

identified 53 articles for inclusion, which increased to 58 articles after we added five articles that we found during reference hand-searching and via feedback from our expert consultants (**Figure 2**).

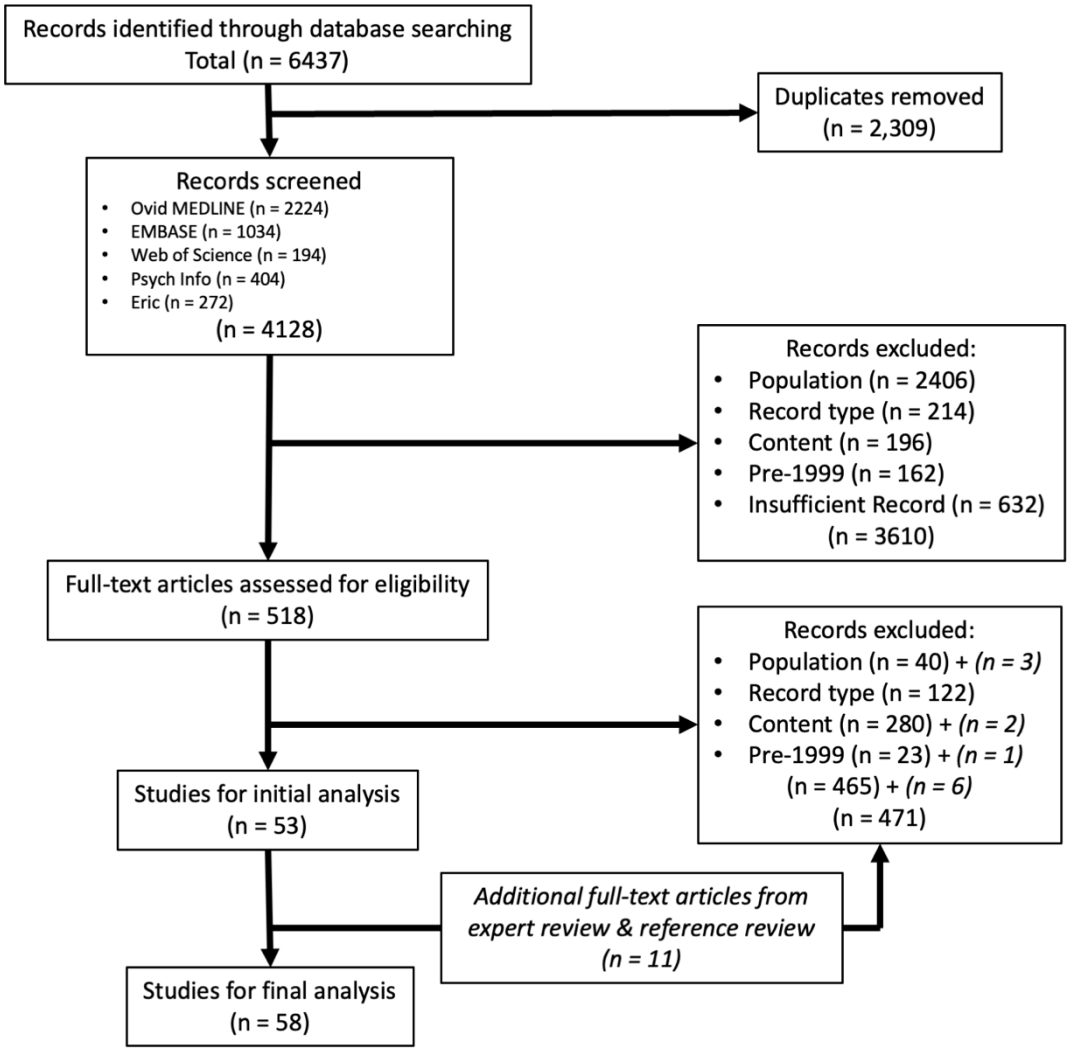

Figure 2. This PRISMA diagram shows the selection process for included articles with articles added from the expert review in *italics*.

Table 1 presents a summary of the results with more detailed results available in appendix 3. All of the articles were published in peer-reviewed journals with about half (n=28, 48%) appearing in educationally focused journals, such as *Academic Medicine, Academic*

Psychiatry, Advances in Health Sciences Education, BMC Medical Education, Journal of Continuing Education in the Health Professions, Journal of Graduate Medical Education, Journal of Surgical Education, Medical Education, and *Medical Teacher*. The remaining articles were published in more clinically focused journals such as *Annals of Surgery, American Journal of Roentology, American Journal of Surgery*, or *Gastrointestinal Endoscopy*. Most studies involved surgical specialties (n=21, 36%), mostly general surgery (n=18, 31%), and general internal medicine (n=14, 24%) training programs with half of the study locations in the United States (n=29, 50%), half outside the United States (n=29, 50%), and one study including locations in both the United States and Canada. One study did not report the location. Among the studies outside the United States, the most common locations were Canada, Netherlands, India, and the United Kingdom. The motivation for most studies was, at least in part, driven by external accreditation bodies (n=37, 64%).

The study populations were generally small (Median=30 subjects, Range 1-216). Most participants were in GME (Median=100%, Range 14.8-100%). Female trainees, among studies that reported sex (n=25, 43%), represented about half the population (Median=46%, Range 22.1-89.1). The study populations contained a mix of training years in most studies (n=47, 81%) with a median 2.5 training years (Range 1-6).

The study quality was variable. The majority of the studies were quantitative (n=46, 79%) with a comparison group in less than half (n=22, 38%) and randomization in less than one-quarter (n=13, 22%). Most studies were conducted at a single institution (n=43, 74%) and in an urban setting (n=53, 91%). Almost all studies had an explicitly stated study purpose (n=55, 95%). Only seven studies (12%) provided a definition for self-monitoring or the self-monitoring

equivalent term used. Definitions included: “in-the-moment awareness of how learning is progressing” (Brydges et al., 2016), “thinking about thinking” (Bond et al., 2004), and “the ability to attend to one’s actions as well as to their effects … mindful attention to one’s thoughts and emotions” (Könings et al., 2016). The quantitative studies and quantitative components of mixed-methods studies earned a mean score of 69% (Range 39-97%) of the available points on the MERSQI. The qualitative studies and qualitative components of mixed-methods studies earned a mean score of 70% (Range 48-93%) of the available points on the SRQR.

Self-monitoring has been conceptualized in the literature as a component of SRL. However, only 17 studies (29%) mentioned any type of learning theory or conceptual framework. Of these, five mentioned a variation of SRL (Archer, 2010; Brydges et al., 2010; Cleary & Sandars, 2011; Hiemstra & Brockett, 2012; Koriat, 1997; Nelson, 1990; Sweller, 1988). Other learning theories or conceptual frameworks cited included: reflective-impulsive model of social behavior (Wouda & van de Wiel, 2014), dual processing theory (Surry et al., 2017), adult learning theory (Ravitz et al., 2013), social cognitive learning theory (Bennett-Levy, 2006), sociocultural learning theory (Sargeant et al., 2010), experiential learning theory (Kolb, 1984), creative problem processes (Francis et al., 2001), learner centered model (Wolpaw et al., 2003), social cognitive learning (Bandura, 1977), conscious competence learning model (Maslow et al., 1987), deliberate practice and mastery based learning (Cook et al., 2013; McGaghie, 2008; McGaghie et al., 2010), and self-motives model of feedback (Crommelinck & Anseel, 2013).

The terminology used to describe self-monitoring varied across studies, and most studies focused on procedural skills (n=23, 40%) with self-monitoring occurring post-task. Many

terms were used to mean self-monitoring, including: certainty, cognitive forcing, confidence, debriefing, judgment, metacognition, reflection, reflection-in-action, review, self-assessment, self-audit, self-awareness, self-drive, self-evaluation, self-observation, self-prediction, self-recording, self-reflection, and self-reporting. Self-assessment was used the most frequently (n=32, 55%). The term "self-monitoring" was used in only five studies (9%). The accuracy of self-monitoring was measured in 41 studies (71%), characterized in 25 studies (43%), encouraged in 12 studies (21%), and taught in 13 studies (22%). Similarly, most studies focused on self-judgment (n=48, 83%), but many also included self-observation (n=28, 48%) and/or self-reaction (n=25, 43%). The majority of self-monitoring was conducted post-task (n=50, 86%) with most being concurrent (n=45, 78%). Most studies were conducted in simulated environments (n=24, 41%) and most focused on procedural skills (n=23, 40%).

	Number of Studies (Percentage)
Publication characteristics	
Education focused journal	28 (48%)
Quantitative	46 (79%)
Qualitative	6 (10%)
Mixed	6 (10%)
Study Design	
Comparison Group	22 (38%)
Randomized	13 (22%)
Single institution	43 (74%)
Study population	
Single training year	7 (12%)
US Training location	29 (50%)
Internal Medicine	14 (24%)
Internal Medicine Subspecialties	6 (10%)
Surgical specialties	21 (36%)

Other	18 (31%)
Quality Markers	
Purpose stated	55 (95%)
Learning theory or framework stated	17 (29%)
Definition of self-monitoring equivalent	7 (12%)
Task Setting	
Inpatient	2 (3%)
Operating room	12 (21%)
Outpatient	12 (21%)
Simulation	24 (41%)
Other	8 (14%)
Task characteristics	
Communication	11 (19%)
Interpretation	14 (24%)
Management	15 (26%)
Procedural	23 (40%)
Team based	7 (12%)
Timing of self-monitoring	
Pre-task	3 (5%)
During task	10 (17%)
Post-task	50 (86%)
Concurrent	45 (78%)
Retrospective	21 (36%)
Role of self-monitoring in study	
Characterized	25 (43%)
Encouraged	12 (21%)
Taught	13 (22%)
Measured	41 (71%)
Type of self-monitoring	
Self-judgment	48 (83%)
Self-observation	28 (48%)
Self-reaction	25 (43%)

Table 1. Shows the characteristics of the collection of the included articles.

Discussion

Exploring the role of self-monitoring in GME through a scoping review was challenging due to the heterogeneity of the literature, lack of clear definitions, and the blurry continuum between self-monitoring and self-assessment. While we attempted to provide clear definitions delineating self-monitoring and self-assessment during the introduction and for the purposes of the inclusion criteria, the reality of how these terms are used in the literature is much more opaque. Nonetheless, we suggest that the concepts of self-monitoring and self-assessment lie on a continuum (**Figure 3**). Some moments of self-awareness are very clearly *self-monitoring*; whereas others are *self-assessment*; and still others, somewhere in the middle. We propose that the continuum lies along the coordinates of *time* and *context*. Time measures the temporal proximity of the self-awareness and actual or relived (e.g., video observed) task performance. When self-awareness occurs without temporal separation from the task, the self-awareness is time-dependent. Context measures the specificity of self-awareness to a single performance of a task. When self-awareness occurs about a specific, single performance of a task, the self-awareness is context-dependent. When self-awareness is both time-dependent and context-dependent, then it is *self-monitoring*. Self-awareness that fails to meet both context and time dependence is *self-assessment* (**Figure 3**). However, not all self-assessment is equivalent. Self-assessment that achieves context dependence is more like self-monitoring than self-assessment that lacks context dependence. And, perhaps, some of our findings in this review apply to context-specific self-assessment. Self-awareness that is time-dependent and context-independent is unlikely to occur, given a person would have to multitask to do this –

performing one thing, while thinking about another (Skaugset et al., 2016). We have labelled this type of self-awareness as undefined self-awareness.

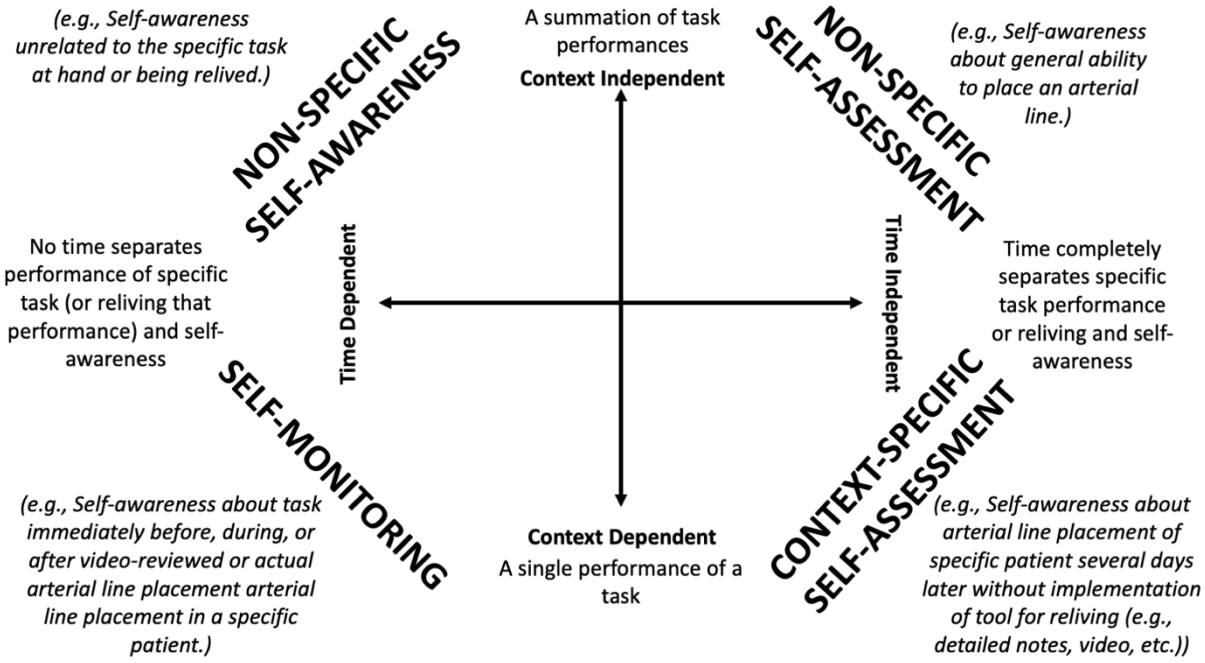

Figure 3. Shows the continuum of self-assessment and self-monitoring based on timing and context. Examples appear in *italics*.

Self-monitoring can look very different depending on the nature of the activity or task performed. During our review of the literature, we encountered self-monitoring that occurred before, during, and following a task. Additionally, self-monitoring occurred concurrently, at the same time as the task performance, or retrospectively, through a method of reliving the task performance. Retrospective self-monitoring often used technology, such as video recording, to bring a person back into mindset of the task (**Figure 4)** in an effort to relive it. The comparison of self-monitoring among studies is difficult because self-monitoring data can be *collected* at many different time points: concurrently before, during, or following task and retrospectively before, during, or following task. At the very least, we encourage investigators of future studies

to clearly describe the methods they used to stimulate and/or record self-monitoring. Methods should include a description of the relationship between self-monitoring and the task – before, during, or following – and the concurrence with the task – concurrent or retrospective. Increased consistency of language will help to clarify the optimal timing of self-monitoring and/or utility of self-monitoring at different moments. Consistent language will make comparing and reproducing studies easier and, as a result, create a more impactful body of literature (Picho et al., 2016). For example, we imagine that the advantages and disadvantages of concurrent and retrospective self-monitoring are different. Some studies suggest that self-judgment improves when performed as retrospective self-monitoring (Jamshidi et al., 2009).

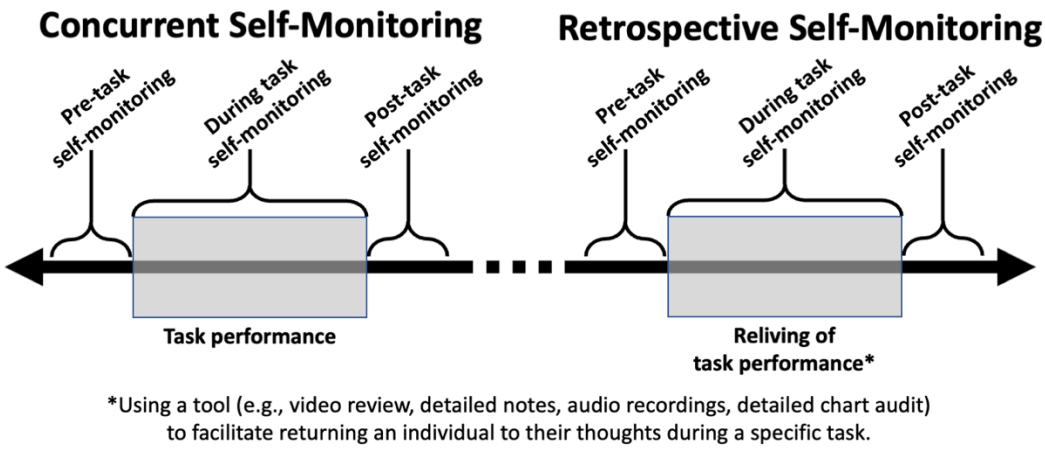

Figure 4. Shows the different times when self-monitoring can occur, providing language to describe the self-monitoring relative to task.

The lack of clear definitions in the literature further contributes to the messiness of differentiating among self-assessment, self-monitoring, and other related terms. We found over 20 terms used to describe the concept of self-monitoring, yet fewer than 20% of studies included a definition of the term used. It is no surprise that the literature is replete with various

terms, given the lack of clear, consistent definitions in the literature. That said, within the health professions education literature, self-assessment and self-monitoring *have been* clearly defined (Eva & Regehr, 2009, 2011). We hypothesize, then, that the lack of definitions may be related to the wide variety of journals that publish, and investigators who examine, self-monitoring. Of the articles that included definitions (n=7), over 85% (n=6) came from health professions education focused journals; only one article emerged from a clinically focused journal. Still, most educationally focused journal articles (78%, n=21) also did not include a definition. Many studies cited accreditation bodies that may have influenced the definitions or lack thereof. Educational accreditation bodies are cited from as early as the late 1990s. While we could not find all the older versions of accreditation standards referenced in the included articles of our scoping review, close examination of current standards from the same accreditation agencies demonstrate a lack of clearly defined terminology. The ACGME Milestones guidebook, for example, mentions “self-assessment” or “self-direct” 16 times without a single definition (Edgar et al., 2020). It is no wonder, then, that clinicians following educational guidance of their accreditation bodies are generally not using consistent language or definitions either. To advance this line of research, we strongly recommend clear definitions that build off prior work. Clear definitions would improve searchability and allow for comparisons, and we urge accreditation bodies to also use clear definitions to drive more effective future research.

Alternatively, perhaps the ubiquitous appeal of self-monitoring drives the heterogeneity of the literature both with respect to self-monitoring implementation and definitions. Self-monitoring is rooted within educational and psychological theories of self-directed learning and

SRL, respectively (Artino et al., 2015; Cleary et al., 2013; Houten-schat et al., 2018). Self-monitoring is a core feature of these theories and some of this work can be quite esoteric. Yet, the appeal of self-monitoring is broad, as evidenced by the fact that over 50% (n=31) of the publications included in our review came from non-educationally focused journals. We wonder: could this finding be related to changes in the language of accreditation bodies? Educationally focused accreditation bodies are cited as a primary motivation for studying self-monitoring as early as the late 1990s (Hildebrand et al., 2009; Wu et al., 2009). And specialty-specific accreditation bodies appeared to motivate researchers in the late 2000s (Scaffidi et al., 2018; Veaudor et al., 2018; Vyasa et al., 2017). Or perhaps the appearance of the concept of self-monitoring in popular books, such as Malcom Gladwell's *Outliers* (2008) or Anders Ericsson's *Peak* (2016), explain some of the ubiquitous interest. Maybe still, the intuitive nature of self-improvement makes self-monitoring feel accessible to almost everyone for a given task. Regardless of the reason, the concept of self-monitoring is being thought about both by educational and psychological theorists and by clinically minded practitioners alike. Publications exploring self-monitoring appear to be rising. Over 40% of our included articles were published within the last 4 years included in our search. While the exact reason remains unclear, the importance of cultivating consistency within the literature seems more important than ever, especially if we want ongoing research to be as productive and meaningful as possible.

Future research should explore gaps in the existing literature. The current literature focuses on post-task self-judgments (41, n=71%) of performance on procedural skills, management, or interpretation. Yet, pre-task and during action self-judgment would seem most critical for patient safety during procedural skills and/or decisions around management and

interpretation. A self-judgment of poor performance after the action will not help the patient under-the-knife or with their discharge paperwork in hand, even if it may improve future practice. Additionally, while the after action self-judgment of procedural skills suggests decent accuracy (Bonrath et al., 2015; Casswell et al., 2016; De Blacam et al., 2012; Ganni et al., 2017, 2018; Jamshidi et al., 2009; Mandel et al., 2005; Osborne et al., 2014; Quick et al., 2017; Scaffidi et al., 2018; Trajkovski et al., 2012; Veaudor et al., 2018; Vyasa et al., 2017; Ward et al., 2003), findings within a procedural context do not uniformly translate to non-procedural contexts. Beyond the accuracy of self-monitoring, several studies actually showed improved procedural skills (Bonrath et al., 2015; Jamshidi et al., 2009). Future research in the procedural realm should build off the demonstrated accuracy of self-monitoring, explore the optimal timing of self-monitoring, and study the positive impact of self-monitoring on clinical practice.

The usefulness of self-monitoring seems promising even in non-procedural tasks. Studies of radiologists and endoscopists have demonstrated improved interpretation with iterative self-monitoring and feedback (Allen et al., 2018; Rosenkrantz et al., 2017; Rzouq et al., 2015). With respect to management, internal medicine residents who engaged in assigned and structured, delayed self-judgment and self-reaction of their management performance improved their diabetic care management for future patients, actually impacting patient outcomes (Holmboe et al., 2005). Regular engagement in a structured, taught form of self-monitoring of physician-patient communication led to increased, self-reported use of best-communication practices (Graddy et al., 2018). In fact, over 65% (n=38) of the studies included in our review showed some improvement or positive correlation with the measured outcome and only 8% (n=5) had a negative impact or correlation. This should provide encouragement

that self-monitoring is worth studying and better understanding. Among the studies with positive findings, self-monitoring was more often paired with coaching or expert feedback (Boet et al., 2011; Bonrath et al., 2015; Bounds et al., 2013; Jamshidi et al., 2009; Kim, 2002; Könings et al., 2016; Plant et al., 2013; Ravitz et al., 2013; Ularntinon & Friedberg, 2016; Wouda & van de Wiel, 2014). None of the studies with negative findings paired self-monitoring with coaching or expert feedback. The observation of the synergism of coaching and feedback with self-monitoring aligns with SRL theory, which describes the interplay among goal setting, motivation, self-monitoring, and cyclical feedback. These observations should influence the development of future studies and help to move the literature away from pure measurement of the accuracy of self-judgment and toward an examination of the methods that encourage self-monitoring and, in particular, the impact of increased self-monitoring of all types (self-observation, self-judgment, and self-reaction) on clinical practice in the context of the other core features of SRL theory. In sum, future research should do a better job utilizing various learning theory frameworks to support study design.

Our results and discussion should be interpreted with a number of important limitations in mind. Most notably, the heterogeneity of the terminology used in the literature on self-monitoring made searching very challenging. As a result, we relied on manual inclusion and exclusion of the literature after a relatively broad search. Manual review increases the probability of human error. We attempted to combat this with secondary reviews of excluded articles, handsearching of the references of our included articles, and leveraging of expert review. The extensive manual review also delayed the completion of our review. Additionally, the quality of research on self-monitoring is highly variable, and much of it is of low quality,

relying on instruments with incomplete or inadequate validity evidence. To address this limitation, we chose transparency, including quality metrics for each included article, regardless of the methodology employed.

Conclusion

Self-monitoring is a complex yet appealing phenomenon that seems promising as a research focus to improve the quality of GME specifically and the safety of clinical practice more generally. The current literature is riddled with challenges, most notably a lack of clear definitions and terminology. We hope this scoping review will establish a foundation for future research on self-monitoring in GME by providing recommendations for terminology, definitions, and theoretical frameworks to guide research questions and study designs.

Chapter 3: Using the think aloud in health professions education: an interview method for exploring thought processes

Authors

W Rainey Johnson, MD, MEd, Anthony R Artino, Jr., PhD, Steven J Durning, MD, PhD

Acknowledgements

A special thank you to Dr. Lara Varpio, who assisted with the literature review, outline, and focus of this manuscript. And thank you to Dr. John Atwood for serving as a participant in the think aloud interview included as an example in this Guide.

Abstract

The think aloud protocol (TAP) has two components, the think aloud interview, a technique for verbal data collection, and protocol analysis, a technique for predicting and analyzing verbal data. TAP is a useful method for those attempting to observe, explore, and understand individuals' thoughts, which remain among the most difficult research areas in health professions education. Notably, the long, complex history and heterogeneous implementation of variations of TAP can make it difficult to understand and implement rigorously. In this Guide, we define the TAP and related concepts, describe the origins, outline applications, offer a detailed roadmap for rigorous implementation as a technique for data collection and/or data analysis, and suggest opportunities for adaptation of the traditional TAP. We aim to arm researchers with the tools to implement a rigorous think aloud interview, while explaining its origins to empower them to adapt the traditional TAP intentionally and justifiably to modern health professions education research.

Introduction

Interviews are a dominant data collection method used by researchers to explore and document the thought processes of individuals. Many different interview approaches have been developed to facilitate that exploration and documentation (Beatty & Willis, 2007; Boyce & Neale, 2006; Corbin & Morse, 2003; Patton, 2015, pp. 432–443), most of which rely on a question-and-response format–i.e., an interviewer asks questions to which the interviewee responds. They also are frequently conducted out-of-context; interview best-practices often recommend finding a private space where the conversation will not be interrupted (Edwards & Holland, 2013). To reply to interview questions, participants engage in reflection and recollection. However, sometimes the data researchers require for their investigations are individuals' real-time narrations of authentic actions. For instance, to understand how assessment guidelines and protocols inform decision making practices, listening to an assessor think through their reasoning can be particularly informative. To capture these narrations, researchers ideally want methods that support *in situ* data collection that happens as close to real-time as possible. One such method is the think aloud interview.

A think aloud is an approach for conducting one-on-one interviews that aims to capture an individual's internal thought processes with minimal intervention from the interviewer (e.g., the interviewer only asks the participant to "think aloud' if no verbalization occurs after a set period of time). A think aloud interview offers researchers a unique means for gathering and recording the thought processes of research participants. The long, and at times complex history of the think aloud makes understanding and using this interview method challenging. Many different terms have been used interchangeably to label the think aloud technique–e.g.,

think aloud method, think aloud interview, think aloud protocol, the think back, think aloud protocol analysis–and the descriptions of the method are at best inconsistent, and at worst contradictory. In fact, the term "think aloud" has been used to define a means of data collection and an approach to data analysis. If the health professions education (HPE) community hopes to harness the power of the think aloud interview, more clarity and specific guidance is sorely needed. In this Guide, we aim to provide such clarity by answering the following questions:

I. What is the definition of a think aloud? And what are the paradigmatic roots?
II. What are the key components of the think aloud interview?
III. When is a think aloud interview useful?
IV. How should the think aloud interview be implemented?
V. How should the rigor of the think aloud interview be assessed? What are the advantages and disadvantages of the method?

I. Definitions and Paradigmatic Roots

The term *think aloud* has historically had a broad scope of inclusion: the term think aloud has been used as a label for a broad combination of data collection approaches that ask research participants to vocalize their thought processes during or immediately after a task. There are three important terms to define and distinguish at the very outset: think aloud interview, protocol analysis, and think aloud protocol (TAP). For our purposes, the *think aloud interview* is the process of collecting data by recording a participant's dialogue of their thoughts with minimal interruption or influence from an interviewer (**Figure 1**). For example, a person

asked to perform a think aloud to name the seventh letter of the alphabet might sing, “A, B, C, D, E, F, G,” while simultaneously making tally marks on a sheet of paper, and then repeat “G.” *Protocol analysis* is a technique for data analysis: a process of creating *a priori* coding for interpreting verbal data and then interpreting verbal data. For example, for the task of finding a specific letter in the alphabet, we might predict that a person will say, “A, 1, B, 2, C, 3, D, 4 …” until reaching the desired letter. However, after interpreting the verbal data, we would revise our prediction as necessary to align with the collected data. In doing so, we would use the think aloud interview to test our hypothesis. *TAP* is this combination of the think aloud interview and protocol analysis.

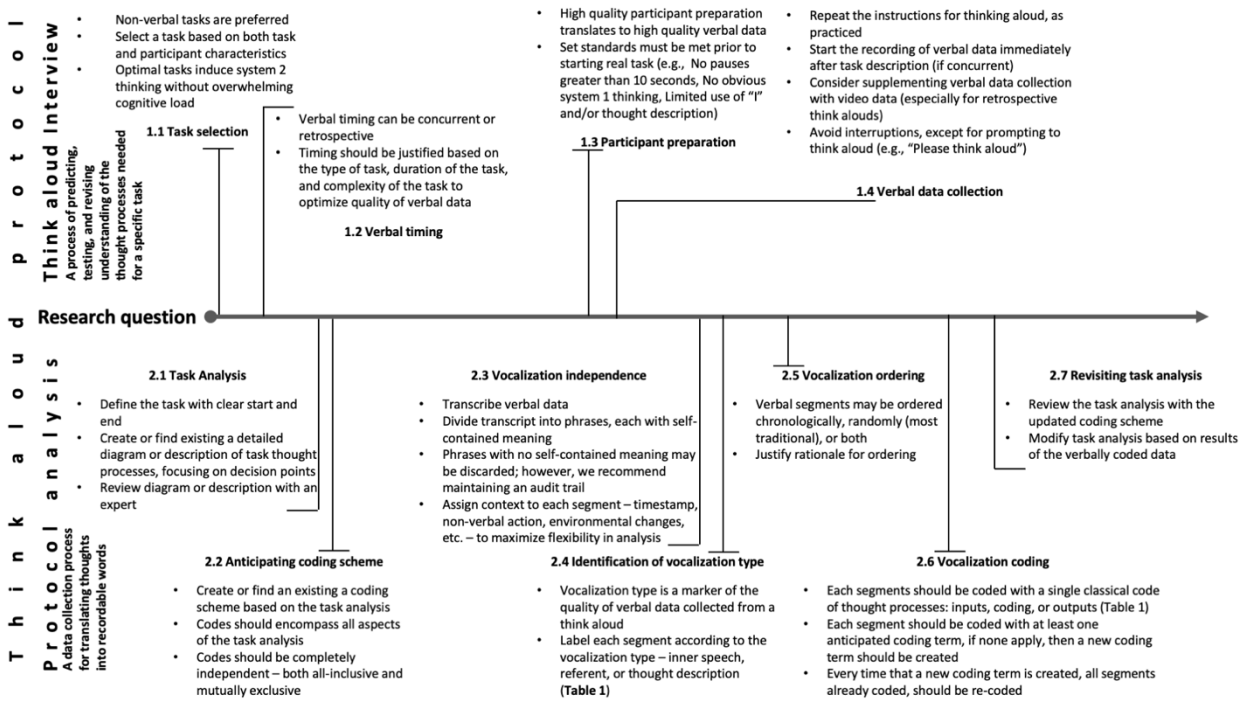

Figure 1. Maps the interrelated timelines of the components of the think aloud protocol – the think aloud interview and protocol analysis – onto a timeline to show the concurrent nature of the individual components with detailed descriptions of each of the steps.

Derived from information processing theory, the think aloud interview transforms working memory thoughts, that is small amounts of readily available, conscious information, into vocalizations, such that the order of vocalization follows the order of thought (Ericsson, 2002; Ericsson et al., 2018). Ericsson and Simon described the think aloud interview and protocol analysis in two separate works from 1980 and 1981. In 1984, they integrated the data collection (i.e., think aloud interview) and analysis (i.e., protocol analysis) in their description of the think aloud protocol (Ericsson, 2002; Ericsson & Simon, 1980, 1981). Protocol analysis uses a pre-specified coding scheme to encode the verbal data and map the thought processes. In other words, once the researchers make a prediction about the thought processes for a specific task, they will create a coding scheme, where each code represents a thought process in their prediction. For our alphabet example, the only code may be "match alpha-to-numeric characters". This code would then be applied to the actual verbal data and, only if it did not represent all the thought processes of the verbal data, would additional codes be added.

In HPE, we often want to know the internal thought processes of learners or even experts to understand how they interpret data or make decisions. The challenge, however, is getting access to that internal thinking and capturing the unobservable. Psychologists who focused on introspection laid the foundation for the think aloud interview (Titchener, 1912). Introspection focuses on individuals' descriptions of their thoughts. However, the consistent misalignment of a research subject's verbalized introspection and observed behavior, undermined the reliability of introspection as a methodology for understanding thought, consistent with the scientific community's preoccupation with objectivity and behaviorism in the growing positivist era (e.g., scientific method to discover immutable, objective facts)

(Ericsson & Simon, 1980; Fuller, 2001). Concurrently, in 1920, Watson wrote about a potentially more objective methodology for researching thought, which relied on asking research participants to simply verbalize their thoughts (Watson, 2009).

The think aloud protocol was born out of the belief that there is an immutable truth that can be best understood through scientific empiricism and hypothesis testing. The positivist orientation to understanding human thinking needed a method for collecting data on that "inner speech"--i.e., verbalized thoughts rather than thought descriptions. Simon described an early version of the TAP, which he further refined with Ericsson in 1980 (Ericsson & Simon, 1980; Simon & Newell, 1971). Their seminal work has remained the cornerstone of TAP for the last 50 years (Ericsson, 2002; Ericsson et al., 2018; Ericsson & Simon, 1981). They described the inseparability of the data collection and analysis, which was likely a byproduct of the historical development of TAP with the need for objectivity.

II. Key components to think aloud protocol

The terminology describing and representing aspects of the TAP can be confusing. Therefore, we will define the main components of TAP and the key terminology needed to understand TAP (**Table 1**).

Term	Definition	Relevance to think aloud protocol
Think aloud interview	The process of collecting data through a participant's dialogue of their thoughts with minimal interruption or influence from an interviewer	The data collection portion of the think aloud protocol
Protocol analysis	A technique for analysis of verbal data that involves anticipating what the verbal data will show and then modifying it based off interpreting verbal data	The most traditional data analysis for verbal data collected with the think aloud interview
Think aloud protocol	The combination of the think aloud interview and protocol analysis	

Positivism	A paradigm, or theoretical belief, that asserts the existence of an immutable truth and prizes objectivity	Laid the foundation for the development of the think aloud protocol that applies hypothesis driven scientific method to evaluating thought processes
Working memory	The retention of small amounts of information that is readily available for execution of cognitive tasks	To vocalize thought, the thoughts must enter working memory to be converted to speech, producing a think aloud
Cognitive load	The portion of the total capacity of working memory used at any given time	Thinking aloud requires working memory adding to the cognitive load of a task.
Intrinsic cognitive load	The mental effort associated with the task and the task alone	The intrinsic cognitive load required to perform the task cannot overwhelm working memory and is an important consideration for task selection.
Extraneous cognitive load	The mental effort associated with everything else besides the task	Thinking aloud adds to extrinsic cognitive load and should be considered in selection of task (i.e., intrinsic cognitive load) and setting (i.e., additional extrinsic cognitive load)
System 1 thinking	Essentially subconscious thinking that occurs quickly and automatically, often through (heuristic) pattern recognition	System 1 thinking is often difficult to articulate, even during the think aloud interview, being re-coded in working memory as thought descriptions, rather than inner thought.
System 2 thinking	Effortful thinking that occupies working memory that occurs deliberately and methodically	Tasks that promote system 2 thinking are ideal for the think aloud interview.
Inner thought (i.e., level 1 vocalization)	Pure verbalized working memory thought.	The type of verbal data that the think aloud interview is designed to capture when most effective.
Referent (i.e., level 2 vocalization)	These are statements of reference, requiring context. For example, referring back to something already discussed or something within the environment.	A common, often contextual, type of verbal data collected during think aloud interview.
Thought description or introspection (i.e., level 3 vocalization)	This is a description of a thought process. Almost like teaching, rather than verbalizing actual thoughts. The descriptions are guided by thought and re-coded as something more intelligible.	A type of verbal data that the think aloud interview is designed to prevent for being collected.
Input	Information that enters working memory	A classical coding term that describes the first step in any thought process and should be coded as part of first layer coding during protocol analysis.
Coding	Interpretation or manipulation of information	A classical coding term that describes the pathway connecting input and output during any thought process and should be coded as part of first layer coding during protocol analysis.
Output	A decision or action	A classical coding term that describes the end-product of any thought process and should be coded as part of first layer coding during protocol analysis.
Reactivity	The notion that the process of vocalizing changes the behavior of the research participant	This is cited as a potential limitation of the think aloud interview, speaking to the importance of minimal interruption during the think aloud interview.
Non-veridicality	The notional that thought processes are unintentionally omitted during a think aloud	This is cited as a potential limitation of the think aloud interview, speaking to the importance of expert involvement in task analysis.

Table 1. Defines terms that are necessary for understanding think aloud protocol.

The think aloud protocol can be divided into the think aloud interview and protocol analysis. The think aloud interview is the approach for collecting thought as verbal data – an interview process for collecting data. It is the process of transforming thoughts from working memory to verbal data, where working memory "is the retention of a small amount of information in a readily accessible form" (Cowan, 2014). The transformation can occur during action, where a participant verbalizes working memory while performing a task. This is called a *concurrent think aloud interview*. Alternatively, the transformation of thought to verbal data can occur immediately after the task completion, via recall. This is called a *retrospective think aloud interview.*

Protocol analysis is an approach for analyzing verbal data meant to represent thought, such as data collected using the think aloud interview. Protocol analysis tests a hypothesis or set of hypotheses about thought processes using verbal data collection and an *a priori* (or pre-data collection) coding scheme. The goal is to transform verbal data into quantifiable data.

The collection of rigorous data using the think aloud interview is arguably sensitive to the participant's cognitive load. In other words, the mental effort that the task requires likely impacts the quality of the think aloud data. Cognitive load is the portion of the total capacity of working memory used at any given time. Cognitive load "is determined by the number of information elements that need to be processed simultaneously within a certain amount of time" (Leppink et al., 2015). Approximately four novel information elements can be held in working memory at a time, unless the information can be chunked, where multiple information

elements are represented by a single chunk (Leppink et al., 2015). For example, remembering the individual digits of a 10-digit phone number is challenging; however, it is made feasible by *chunking* the phone number into three parts (i.e., 3 digits - 3 digits - 4 digits). Participating in a think aloud interview uses up some of the participant's cognitive capacity, thereby increasing cognitive load, which is why it is an important consideration for task and setting selection.

Intrinsic and extraneous cognitive load are two relevant components of cognitive load to the think aloud interview. Intrinsic cognitive load is the mental effort associated with the task and the task alone. Alternatively, extraneous cognitive load is the mental effort associated with everything else besides the task. For example, if a person is asked to fold an 8.5" x 11" piece of printer paper into a square, the intrinsic cognitive load is the working memory required to plan, execute, and complete the task. If the same person is asked to complete a think aloud interview during the task, then the verbalization of her working memory is an extraneous cognitive load. If two colleagues are outside the door having an argument, tuning them out adds extraneous cognitive load. Minimizing extraneous cognitive load during a think aloud interview is critical for optimizing data quality.

The task type and the experience of the person completing the task influences the intrinsic cognitive load. For example, an expert musician asked to name a music note will respond automatically, while a novice asked to name the same music note may count the bars or employ a memory device, such as "Every Good Boy Deserves Fudge" to march through the notes of the treble clef, effortfully working toward the same answer that was automatic for the expert musician. The expert musician uses system 1 thinking, thinking that occurs quickly and automatically, often through pattern recognition. The novice uses system 2 thinking, an

effortful thinking that occupies working memory and that occurs deliberately and methodically. The type of system thinking that a task evokes for a particular participant influences the detail of the verbal data collected during a think aloud interview.

III. When to use think aloud protocols

The think aloud protocol has been used in many different fields, including engineering, computer science, sport, and health professions, to name just a few (Durning et al., 2013; Duschl et al., 2015; Göransson et al., 2007; Kelley et al., 2015). The research question helps dictate the type of think aloud interview implemented based on the task type, setting, and complexity. We will focus on examples of potential and actual applications of TAP within HPE research with descriptions of rationale for implementing concurrent versus retrospective think aloud interviews.

Traditionally, TAP has been used to verify predicted thought processes for a specific task. For example, the think aloud interview has been used to optimize electronic health record usability. Richardson et al used TAP to improve the ease of use of an embedded clinical prediction tool and suggested order sets (Richardson et al., 2018). Similar studies have been conducted to verify the usability of clinical decision support systems (Kilsdonk, 2016). Both studies used concurrent think aloud interviews, given that the task of interacting with the electronic health record and clinical support tools did not require verbalization. Concurrent think aloud interviews are ideal for collecting verbal data on the thought processes of non-verbal tasks like placing a central or arterial line, performing a laparoscopic cholecystectomy, or injecting corticosteroids into a joint space. Similar verbal data collection could be used to

understand the thought processes for visual or auditory interpretations, such as a pathologic review of a blood smear, inspection of the retina with a slit lamp, or interpretation of a chest x-ray.

The think aloud protocol could also be useful for exploring thought processes for a specific task and to identify extraneous versus intrinsic cognitive load. For example, Kilsdonk et al (2016) conducted an iterative TAP. That is, they repeated TAP. In the first iteration, they explored the clinical decision-making thought process and then used this to develop a clinical decision support system. Then, they verified that the clinical decision support system usability was optimized with a TAP (Kilsdonk, 2016). An iterative TAP can be particularly useful for refining a tool for clinical practice or even a practice environment. The iterative process could be used to optimize an electronic medical record for pre-rounding on a series of outpatients or inpatients. Imagine a nurse working in a busy family medicine practice, an iterative TAP may be used to refine the practice space and facilitate switching among patients with different needs – children, pregnant mothers, or frail older adults. An iterative TAP may include concurrent or retrospective think aloud interviews.

In some instances, investigators have justified a combination of concurrent and retrospective think aloud interviews, such as for exploring the clinical reasoning of pharmacists as they process prescriptions (Croft et al., 2018). By complementing the concurrent think aloud interview conducted during a simulated patient encounter, the investigators captured additional clinical reasoning that may have been lost due to task complexity or task type. In this case, the simulated patient communication, a verbal task, would have compromised the

participants' ability to concurrently vocalize their thoughts. Similar studies have been conducted on clinical reasoning during simulation with nurses (Burbach et al., 2015).

Retrospective think aloud interviews are helpful for tasks that involve verbal communication, and/or have such high complexity or acuity that concurrent vocalization could be compromised, or when vocalization might reduce data quality of another measurement. For example, Durning et al (2012 & 2013) evaluated clinical reasoning of internal medicine physicians on multiple choice assessments using functional magnetic resonance imaging and a retrospective think aloud interview, as jaw motion would have compromised the quality of the fMRI images (Durning et al., 2012, 2013). In another study evaluating the cognitive rationale for applying different motor learning techniques among expert physiotherapist, a video-stimulated retrospective think aloud interview was used because the patient-physiotherapist interaction required verbal communication (Kleynen et al., 2017). Retrospective think aloud interviews could be useful for exploring the rationale for selecting communication tools to break bad news or respond to emotion. Similarly, understanding the clinical decision-making in response to an in-hospital cardiac arrest would require a retrospective think aloud interview, since the act of participating or leading an emergency response precludes participation in a think aloud interview.

The applications of TAP to explore and understand thought processes in health professions are bountiful. In this Guide of TAP, we will further illustrate TAP using an example of clinical interpretation. We selected an expert in the field of cardiology and echocardiography and performed a TAP during image interpretation to explore the clinical reasoning associated with echocardiography interpretation.

IV. A practical approach for implementing the think aloud protocol

The think aloud protocol includes data collection and data analysis. For the sake of clarity, we will describe the think aloud interview and then protocol analysis. It is important to note, however, that protocol analysis is only one method that can be used for data analysis purposes with the think aloud interview. Think aloud verbal data can be paired with other qualitative analysis techniques even though, historically, the think aloud interview and protocol analysis have sometimes been described as inseparable. Most of the think aloud interview occurs temporally before the protocol analysis, but parts of the protocol analysis can occur before any data is collected (**Figure 1**).

The dominant approach – think aloud protocol

The think aloud interview involves four steps (**Figure 1**): (1.1) task selection, (1.2) verbal timing selection, (1.3) participant preparation, and (1.4) verbal data collection.

(1.1) Task selection:

The optimal task for the think aloud interview should allow participants to vocalize at least every 5-15 seconds. The more frequent the vocalizations, the better. The longer the period without vocalization, the higher risk there is for missing a part of the thought process (Fox et al., 2011). Tasks should ideally be cognitive tasks because this allows for concurrent verbalization, that is vocalization of thoughts at the same time as task performance, which is thought to be more reliable (Koro-Ljungberg et al., 2013). Tasks should not result in high

cognitive load because really challenging tasks make it difficult to simultaneously vocalize internal thoughts. It is important to remember that the expertise of the participant may influence the difficulty of the task. For example, if the task is unfamiliar to the participant, it will be more difficult than if it is a task performed regularly. Along the same lines, the task environment should allow for verbalization without easy distraction – quiet environments preferred. Minimizing the extraneous cognitive load is often at odds with executing the task in a real-world or high-fidelity environment. Yet, removing the task from the real-world environment may alter participants' thought processes. For example, the task of using a decision-making tool to determine which patients with shortness of breath get a computed tomography scan for pulmonary embolism may render very different thought processes in a quiet, simulated environment, rather than a chaotic, noisy emergency room department. However, to optimize the decision tool, understanding the thought processes in the emergency department would be essential. We believe that minimizing extraneous cognitive load should not interfere with maintaining the critical situational fidelity of a task.

(1.2) Verbal timing selection:

Verbalization can be concurrent or retrospective. A concurrent think aloud interview is a research participant's vocalization of his or her thought processes *during* the performance of a specific task. Ericsson and Simon argue that concurrent think aloud interviews yield the highest quality verbal report data. A retrospective think aloud interview has research participants vocalize their thought process *immediately after* having completed the specified task (Ericsson, 2002; Fonteyn et al., 1993). If the task requires verbal communication, for example breaking

bad news to a patient, then concurrent verbalization is impossible since the participant cannot have a conversation with a patient and also verbalize her thoughts. When concurrent verbalization is not possible, we suggest recording the encounter and then requesting verbalization during self-observation of the video after completing the actual encounter (Artino et al., 2014). Leveraging tools, such as video or audio recording, may be especially important if the duration of the task is long, where the recall required for retrospective think aloud interview is typically unreliable (Ericsson & Simon, 1981; Koro-Ljungberg et al., 2013). If a task is particularly challenging, then retrospective verbalization may also be more beneficial because the act of verbalizing during the task may impact the task duration and, despite lack of supporting evidence, possibly performance (Eccles & Arsal, 2017; Fox et al., 2011). The selection of timing should, therefore, be justified based on the complexity of the task, such that the timing optimizes the quality of the vocalization data.

(1.3) Participant preparation:

Some participants find verbalizing during a task quite natural, and others do not. The quality of the verbal data depends on the participants' ability to verbalize their thoughts. To optimize quality of data collection, each participant should have clear instructions and an opportunity to practice (**Table 2**) (Eccles & Arsal, 2017; Ericsson et al., 2018; Ericsson & Simon, 1980; Fonteyn et al., 1993; Göransson et al., 2007; van Someren et al., 1994). Instructions should include the expectations for continual verbalization of thoughts and naming of cues that the interviewer will use to remind the participant to verbalize if no verbalizations have been made for more than 5-10 seconds. Then participants should practice verbalizations during a familiar task, such

as completing a word search or making a sandwich. If the participant struggles to verbalize throughout the practice session, then the practice session should be repeated with a different task until the participant is comfortable vocalizing. If iterative practice sessions are not feasible, then some think aloud experts suggest that the participant should be excluded from further participation in the research (van Someren et al., 1994).

Participant friendly description of the think aloud interview	*I am going to ask you to think aloud while performing a task. What I mean by thinking aloud is that I want you to say your thoughts out loud from the moment you finish hearing the instructions for your task until completion. Even if your thoughts wander from the task, I want to hear them. Say as much as you feel comfortable saying. Don't try to plan or explain what you say, just act as if you were speaking to yourself.* If the research calls for a retrospective think aloud, then the language above should be modified and the practice sessions should include practice retrospective think aloud.
Name cue to facilitate thinking aloud	*If you pause for too long without saying your thoughts aloud, then I'll remind you by saying, "please think aloud".*
Practice activities	*We are going to do a warm-up activity to help you feel comfortable with thinking aloud. What questions do you have before we get started with some practice?* Select practice activities that align with plan research task (i.e., non-verbal with non-verbal; verbal with verbal): ○ Name five animals that live at the zoo ○ Fold this 8.5" x 11" paper into a square ○ What is the fifth letter before "M" in the alphabet? ○ Perform word search to find as many words as possible within 2 minutes. ○ Using the items and tools in front of you … ■ Make a cup of tea ■ Make a sandwich ■ Reproduce this (e.g., optical illusion) image
Provide feedback	● Example feedback on task description: *When you said, "I am folding the paper," you were describing your actions.* Select a feedback method below or use a combination: ○ *Let's go back to that moment and I want you to try to say your thoughts.* ○ *Instead, I'd like to hear your thoughts. For example, this might be "match the corners … equal … nope".* (As you model) ● Example feedback on thought description: *When you said, "Zoos organize like animals with like. I am thinking of the last time I went – I say the Big Cats, Ape house, Reptile house …" you were describing your thoughts. You were modifying your thoughts into more coherent and acceptable speech. To the best of your ability, I want to hear exactly what you are thinking without modification – like a stream of consciousness, For example, "Big cats – lion, tiger, cheetah … Africa … Bobcat, Americas … panther, mountain lion, lynx".* ● Example of feedback on system 1 thinking: *You answered that really quickly. You said, "G" and nothing else. How did you get there?* (Pause for answer) *What you've just done is describe your thoughts. I imagine that you came to your answer so quickly because it was automatic for you, almost subconscious. To the extent possible, I really want you to try and say anything that comes to mind.* ● Feedback on pauses: *You paused after saying, "it's a rectangle," what were you thinking?* (Pause for answer) *I want to hear those thoughts. When people get stuck, they often pause their thinking aloud and, for me, these are the most important moments to record. Let's try another task.*

Table 2. Shows an example of practice instructions and activities for a think aloud interview. Example statements by the interviewer are *italicized*.

(1.4) Verbal data collection:

The verbal data collection should start with a reminder of the expectation and cue words, using the same language as the instructions for the practice tasks. The interviewer should then define the task and, if concurrent, vocalization of thought should immediately begin. During the task, the interviewer should minimize interruptions – verbal and non-verbal – strictly following the cues outlined in the expectations and practice session. In other words, the interviewer should only interrupt with the pre-specified cue—e.g., think aloud—when a participant has not vocalized for the preset amount of time (e.g., 5-10 seconds) Any interruption may increase cognitive load and impact data quality (Ericsson & Simon, 1980; Fonteyn et al., 1993). Verbal data should be recorded, while minimizing distraction. Video data may be a helpful adjunct in some instances, especially to record actions and if alternative data analysis methods are used.

Protocol analysis involves (1) task analysis, (2) anticipated coding scheme, (3) vocalization independence, (4) identification of vocalization type, (5) vocalization ordering, (6) vocalization coding, and (7) revisiting task analysis (**Figure 1**).

(2.1) Task analysis:

The task should be defined with a clear and discrete start and endpoint. The research team must create a description or diagram that shows how a task is hypothesized to be completed, focusing on the decision points and how decisions will likely be made. A task expert should review the diagram or complete description of the hypothesized task performance. Alternatively, if an existing task analysis exists, it should be used.

(2.2) Anticipated coding scheme:

Based on the task analysis, the research teams must create a coding scheme. When feasible, using an existing coding scheme may help increase the reliability of the analysis and assist in making comparisons and connections among other research studies easier (Atman & Bursic, 1998; Grubbs et al., 2018). The goal of the coding scheme is to create the minimal number of coding terms to completely represent the task analysis. Every aspect of the task analysis should be captured by the coding scheme. The coding scheme should adhere to the following principles: (1) Related to hypothesis (i.e., task analysis), (2) Create coding categories that are simultaneously all-inclusive and mutually exclusive (Yang, 2003).

(2.3) Vocalization independence or segmenting:

The vocalizations should be transcribed verbatim. Then, the research team must read the transcript and segment the vocalization transcript into discrete segments or phrases, such that each phrase has self-contained meaning and contains participant expression alone (Yang, 2003). In other words, do not include interviewer or environmental cues.

An *a priori* coding scheme is helpful for looking for segments, since the coding scheme provides guidance on what self-contained phrases may be found. There is no need to segment the entire transcript. Where phrases have no self-contained meaning, it is permissible to discard the data (Ericsson et al., 2018; Ericsson & Simon, 1980). However, if in doubt that a phrase may have self-contained meaning or if a phrase may have significant albeit uncertain meaning, we favor

transcribing it as a segment. Additionally, we recommend maintaining an audit trail to record the changes made from raw transcript to verbal segments included in analysis, as to enhance trustworthiness (Carcary, 2009). In the process of segmenting, we suggest pairing the segment with the timestamp or assigning an ordinal value (such that the chronological order could be reconstructed) and describing non-verbal context or actions. This can be done using a simple spreadsheet and may well be ignored but allows for re-coding with greater context if context-free coding, which is the most traditional way to perform protocol analysis, feels incomplete.

(2.4) Vocalization type:

For each vocalization segment, determine the level of vocalization. There are three levels of vocalization: inner speech, referent or label, and thought description (**Table 1**) (Ericsson et al., 2018; Ericsson & Simon, 1980). The think aloud interview strives to capture inner speech to the greatest extent possible.

a. Inner speech: This is thought verbalized through working memory.

b. Referent: These are statements of reference, requiring context. For example, referring to something already discussed or something within the environment.

c. Thought Description: This is a description of a thought process. Almost like teaching, rather than verbalizing actual thoughts. The descriptions are guided by thought and re-coded as something more intelligible.

(2.5) Vocalization ordering:

Protocol analysis purists favor randomly ordering the verbal segments. The goal with random ordering is to remove the context to the greatest extent possible, so that the coding and interpretation is focused exclusively on the thought without influence from other factors, such as a prior segment or contextual clue that may influence the analysis (Yang, 2003). While we respect the potential value of this approach, we also appreciate that context plays an intimate role in learning, self-monitoring, cognitive load, and reasoning. Therefore, we believe that investigators should order verbal segments as they see fit and remark on their influence on the analysis and the rationale for the ordering in disseminated work. One of the reasons that we suggest adding a timestamp for each segment is to allow for random ordered coding and then repeat coding in chronological order. The iterative process may provide more information about thought processes and the dependence of thought processes on the length of task and/or context.

(2.6) Vocalization coding:

The vocalizations should be coded in two layers. In the first layer, each segment should be coded to capture the role within the thought process. The classic terms are input, coding, and output (Ericsson & Simon, 1980; Lundgrén-Laine & Salanterä, 2010). Input describes information that enters working memory. Coding describes the interpretation or manipulation of information. Output describes a decision or action (**Table 1**). In the second layer, the anticipated coding scheme should be applied. Each segment should have at least one anticipated coding label (Ericsson & Simon, 1981; Yang, 2003). Where two or more codes apply, verbal segments should have multiple coding labels assigned. If segments do not fit within the a

priori scheme, new coding terms should be created. As new terms are created, each already coded collection of segments from a single participant should be re-coded.

(2.7) Revisit Task Analysis:

With the coding complete, researchers should revisit the task analysis and adjust based on the coding results. Some researchers may find it helpful to re-order the verbal segments chronologically, even if coded in a random order, as the task analysis is revised.

Application of the dominant approach

We have used TAP to understand how experts interpret transthoracic echocardiograms. We developed a task analysis based on conversations and observations of expert cardiologists (**Appendix 4**). Derived from the task analysis, we created a list of a priori coding terms, felt both mutually exclusive and all-inclusive for the task of interpreting an echocardiogram (**Table 3**). Then, we prepared and performed a think aloud interview, as described above (**Figure 1**).

A priori coding		Added coding	
Code	Definition	Code	Definition
Interpret question	Predict data needed to answer consult question	Apply differential diagnosis	Apply differential diagnosis to data for goodness of fit
Create gestalt	Categorize data as normal or abnormal	Name data	Describe the data being interpreted
Compare	Comparing two or more views, perspectives, or data points from different time periods		
Triangulate	Integrating two or more views, perspectives, or data points from the same time period		
Finalize interpretation	Decision complete and ready for documentation		

Doubt	Expressing uncertainty		
Assess quality	Qualifying the goodness of the data		

Table 3. Shows the original coding scheme based on the task analysis and the added coding schemes derived from the coding process.

We applied protocol analysis to a portion of the recording as an example. In a spreadsheet, we segmented the verbal data into discrete phases with self-contained meaning, removing extraneous vocalizations, while simultaneously maintaining an audit trail (**Table 4**). We assigned the appropriate timestamps and non-verbal data (from observations during the think aloud interview) to each of the verbal segments, creating a spreadsheet of raw data. The timestamps and non-verbal data allowed for the option of returning to the raw data to perform a protocol analysis with greater contextualization. However, to demonstrate the most traditional approach, we iteratively coded the randomly ordered vocal segments, starting with the original coding scheme that evolved as we coded (**Table 3 & 4**). Given that we only conducted a think aloud interview with a single individual, we did not revisit our task analysis.

Transcript	Vocalization Segmenting			Vocalization Coding (Random order)					
The left atrial area during end-systole in both the two and four chamber view looks normal. The wall motion segments appear normal. There is mitral regurgitation that appears a little eccentric over the posterior wall and I'm thinking the inferior wall wall motion is normal so this isn't secondary to restricted posterior leaflet but not much anterior leaflet prolapse either maybe there could be another explanation such as a perforation on the posterior leaflet not certain but I suspect that just a little bit of prolapse.	Time	Vocal Segment	Action	Verbal segment	Level of vocalization	Classical coding	Primary Task analysis coding	Secondary task analysis coding	Type of system thinking
	0:17:41	the left atrial area during end-systole in both the two and four chamber view	Viewing two chamber and four chamber views	the left atrial area during end-systole in both the two and four chamber view	Referent	Input	Name data		N/A
	0:17:49	looks normal	Viewing two chamber and four chamber views	not certain	Inner speech	Coding	Doubt		System 2
	0:17:52	the wall motion segments	Viewing two chamber and four chamber views	such as a perforation on the posterior leaflet	Inner speech	Coding	Triangulate		System 2
	0:17:57	appear normal	Viewing two chamber and four chamber views	I'm thinking the inferior wall wall motion is normal	Thought description				System 1
	0:17:59	there is mitral	Looking at	appears a little eccentric over the posterior wall	Inner speech	Coding	Create gestalt	Doubt	System 2
				appear normal	Inner speech	Coding	Create gestalt		System 1

	regurgitation	mitral valve with color doppler
0:18:06	that appears a little eccentric over the posterior wall	Looking at mitral valve with color doppler
0:18:14	I'm thinking the inferior wall wall motion is normal	Looking at mitral valve with color doppler
0:18:22	so this isn't secondary to restricted posterior leaflet	Looking at mitral valve with color doppler
0:18:28	not much anterior leaflet prolapse either	Looking at mitral valve with color doppler
0:18:34	maybe there could be another explanation	Looking at mitral valve with color doppler
0:18:41	such as a perforation on the posterior leaflet	Looking at mitral valve with color doppler
0:18:45	not certain	Looking at mitral valve with color doppler
0:18:48	I suspect just a little bit of prolapse	Looking at mitral valve with color doppler

the wall motion segments	Referent	Input	Name data		N/A
not much anterior leaflet prolapse either	Inner speech	Coding	Apply differential diagnosis		System 2
there is mitral regurgitation	Inner speech	Output	Finalize interpretation		System 1
looks normal	Inner speech	Coding	Create gestalt		System 1
maybe there could be another explanation	Inner speech	Coding	Doubt	Apply differential diagnosis	System 2
I suspect just a little bit of prolapse	Inner speech	Coding	Apply differential diagnosis	Triangulate	System 2
so this isn't secondary to restricted posterior leaflet	Inner speech	Coding	Apply differential diagnosis	Triangulate	System 2

Table 4. Shows processing of verbal data from raw transcript to verbal segments with contextual information to coding.

An alternative approach to data collected from the think aloud interview

The choice of analysis depends on the research question. While other analyses are less well-established than protocol analysis, we support the notion that alternative analysis techniques may be useful, especially (and not only) for inductive analysis. For example, thematic analysis, a flexible method for analyzing qualitative data, aids researchers in deriving themes to understand behaviors, attitudes, experiences, and thoughts (Kiger & Varpio, 2020). Thematic analysis may be a useful pairing with the think aloud interview to explore factors that influence self-monitoring or uncertainty, understand decisions that lead to unprofessional behaviors, identify triggers for converting from 1 to system 2 thinking, or attitudes that

promote culturally sensitive communication (Artino et al., 2014; Burbach et al., 2015; Moulton et al., 2010; Varpio et al., 2015). While the best available evidence for the TAP is grounded in positivism and post-positivism, we encourage researchers to use the think aloud interview with other paradigms and recommend that researchers explicitly state the underlying paradigm, or theoretical belief, of the analysis applied.

V. Evaluating the rigor of think aloud protocols

Optimizing data quality

The rigor of the think aloud interview depends on the quality of the data collected. While we have enumerated common pitfalls throughout this Guide and in this section, we have summarized them in **Table 5**. Accurate verbal transformation of working memory thoughts is the cornerstone of the method. Attention to cognitive load, adequate preparation, and justifying the timing of thought verbalization are the three key steps to maintaining and/or evaluating the rigor of the think aloud interview.

Common pitfall	Potential manifestations	Suggestions for avoiding and/or preventing pitfall
Poor task selection	● Too much system 1 thinking ● Too many pauses and/or long pauses ● Too much thought description	Trial task Add complexity to task Adjust verbal timing (e.g., concurrent v. retrospective) If retrospective, cue think aloud with video
Poor preparation	● Too many pauses and/or long pauses ● Too much thought description	● Poor preparation ● Poor task selection too complex ● See table 3
Disrupting inner speech	● Frequent interruptions during think method ● Statement other than identified cue ● Conducting another interview type simultaneously ● Task requires vocalization independent of think aloud	● Interviewer should practice ● Interview location should have minimal non-task related interruptions ● Alternative interview types should occur separately, if at all ● Adjust verbal timing (e.g., concurrent v. retrospective) ● If retrospective, cue think aloud with video ● If necessary extraneous cognitive load (e.g., situation or environment essential), then pair verbal data with contextual observation and consider coding in chronological order

Table 5. Recommends methods for avoiding the most common pitfall when executing a think aloud protocol.

Excessive cognitive load makes continual verbalization of working memory thoughts very difficult, increasing the probability of missing periods of working memory thought (Charters, 2003). At the same time, low cognitive load may be associated with tasks that rely predominantly on system 1 thinking, making verbalizing working memory difficult because tasks occur almost automatically. For example, an expert clinician recognizes an abnormal lab value, such as hemoglobin, automatically. Requesting a verbalization of working memory may yield verbalization of thoughts that normally would not occur in real practice. We observed this while our expert cardiologist named anatomy structures and visually assessed ejection fraction. At these moments, he drifted more toward descriptive thought, rather than inner speech.

Optimizing the cognitive load depends on the participant, the task, and her relationship with it. The more familiar the task, the more likely the participant is to implement system 1 thinking and to have chunking strategies to manage more total information in their cognitive load. To optimize the think aloud interview, the task should be appropriate for the participant, complex enough to reasonably elicit system 2 thinking. Alternatively, the less familiar the task, the less complex it should be. If a research question requires an unfamiliar task that is challenging, then breaking up the task into smaller, discrete tasks may help to optimize the quality of the verbal data (Charters, 2003).

Preparing participants can further improve the quality of verbal data. Since verbalizing thought is not normally part of performing (most) tasks, it adds to the extrinsic cognitive load. Practicing the act of verbalizing thought processes helps to minimize the impact of the added

cognitive load. Practicing is essentially a rehearsal for the working memory, helping to make the task of verbalizing thought processes from working memory more automatic. Explicitly defining acceptable performance during a think aloud interview practice session helps to produce high-quality data for desired verbal data collection (Eccles & Arsal, 2017; Göransson et al., 2007). Between practice sessions, we suggest providing feedback to the participant to improve subsequent performances.

Think aloud interviews should be conducted concurrently, whenever possible. Concurrent think aloud interviews more reliably produce verbal data that represent working memory thoughts (Ericsson et al., 2018; Fonteyn et al., 1993; Fox et al., 2011; Koro-Ljungberg et al., 2013). However, some studies suggest that the concurrent think aloud interview does not impact task performance, only duration (Ericsson & Simon, 1984; Henry et al., 1989). If a retrospective think aloud interview must be used, then clear justification should be made. For example, retrospective think aloud interview may be best if the cognitive load of the task is too demanding and cannot be adjusted or if the task requires speech. Retrospective think aloud interview data is likely higher quality if the data is collected as close to the actual task as feasible. So, splitting the task into shorter, discrete tasks may be one method for increasing the quality of retrospective data. Cues that help to bring the participant back to the moment, such as video, may also be helpful (Koro-Ljungberg et al., 2013).

Protocol analysis, with its positivist and post-positivist roots, may not be the most appropriate analysis technique for verbal data collected through the think aloud interview. For example, if the goal is to verify a predicted thought process, then protocol analysis should be used. If the research question focuses on exploration, then alternative analysis may be more

appropriate. We suggest that any deviations from a strict protocol analysis also be justified. For example, if context is very important then randomly ordering verbal segments for analysis may lose valuable data. In this case, we suggest coding verbal segments in chronological order and with consideration of contextual observations. We recommend that researchers clearly describe and justify their methods. Regardless of approach, we deviate from traditional recommendations with respect to inclusion of transcribed verbal data, embracing the think aloud interview as a qualitative research method.

Despite the origin of TAP, we recognize that it is impossible to fully and objectively "know" the thoughts of an individual. The process of encoding thoughts as language is unlikely a perfect one-to-one mapping. We also caution investigators that regardless of how much they attempt to remain invisible within the data collection and data analysis process, this is impossible. Therefore, examples of transformations that the researchers made from verbal data to vocalization coding should be included in the results. Researchers should maintain an audit trail and should consider how they may have influenced the collection of verbal data through task selection, participant practice, interviewer presence, and vocalization timing. Even more so, researchers should embrace reflexivity with respect to the verbal data analysis, considering their influence on the segmentation, multiple layers of coding, and interpretation/application of data to the task. While TAP may have been developed as a quantitative method, the modern think aloud should be viewed as a qualitative method with incorporation of the rigors of qualitative researcher methods (O'Brien et al., 2014).

Critiques leveled against think aloud protocol

The think aloud protocol is not without critics, who have contested that the process of vocalizing changes the behavior of the research participant and, thereby, the thought processes (Gagné & Smith, 1962; Nisbett & Wilson, 1977). This is referred to as reactivity (Russo et al., 1989). However, Ericsson and Simon (1980) have argued that strict adherence to the method with concurrent thinking aloud and minimal interruption of the researcher has little impact on participant performance (Ericsson & Simon, 1980). A meta-analysis comparing the impact of think aloud interview methods showed that the think aloud interview did not impact research subject performance when compared to silent task performance (Fox et al., 2011). The think aloud interview did, however, increase the time to task completion. These results held true for verbal and non-verbal tasks.

The second major critique of the think aloud interview is that thoughts are unintentionally omitted or altered in the processes of vocalizing, which is called *non-veridicality* (Russo et al., 1989). Ericsson and Simon have refuted the claim of non-veridicality largely through task analysis (Ericsson et al., 2018; Ericsson & Simon, 1981; Fox et al., 2011). Through *a priori*, expert developed thought process maps, Ericsson and Simon have mapped concurrent TAPs and found high degree of correlation between the expected and observed, via think aloud interview, thought processes (Ericsson, 2002). A more recent study using concurrent think aloud and silent thinking while answering multiple choice questions and undergoing brain functional magnetic resonance imaging concluded that the think aloud interview likely has little impact on veridicality and further studies are needed for confirmation (Durning et al., 2013).

The think aloud protocol, as Ericsson and Simon designed, also has limitations with respect to data collection and data analysis. With respect to data collection, the think aloud

interview requires a strict, controlled environment, limiting the utility in situated or work-based learning environments (Ericsson & Simon, 1980). The higher cognitive load in a situated environment may increase the non-veridicality, though an increase in omitted or altered thoughts during vocalization. Excessive cognitive load may undermine the cornerstone of the think aloud interview that relies on verbalization of all working memory thoughts as level 1 (i.e., inner speech) or 2 (i.e., referent) vocalizations (Ericsson et al., 2018). On the other end of the spectrum, tasks that are insufficiently complex and familiar are likely to elicit system 1 thinking, which may result in incomplete vocalizations, as participants may not be consciously aware of the specific thought processes that led them to pattern recognition. In addition to task complexity, task duration may also be limiting. The think aloud interview is ideal for short tasks, limiting utility for tasks of long duration (e.g., greater than 50 minutes). Similarly, protocol analysis is of limited utility for analyzing tasks requiring very complex thought. After all, protocol analysis and the anticipated coding schemes are grounded in linear thought processes (Ericsson & Simon, 1981); however, cognitively complex thought processes may generate non-linearity, which cannot be adequately captured with protocol analysis.

Conclusion

The think aloud protocol and, in particular, the think aloud interview, can offer unique insight into working memory thoughts for verifying, exploring, and understanding cognitive processes in HPE. In this Guide, we have attempted to clearly define the key terminology required to understand TAP and to delineate the think aloud interview and protocol analysis as separable components of TAP. We have also encouraged a modernized view of TAP that

permits the separation for the think aloud interview from protocol analysis. Finally, we have enumerated some best practices for adherence to the dominant think aloud interview method, while simultaneously offering suggestions for the evolution of think aloud interview and analysis techniques to help further understand thought in health professions education, even as research questions necessitate crossing paradigms.

Chapter 4: Discussion

The scoping review and methods paper highlight the criticality of clearly defined terminology and combine to offer a path forward for better understanding the process of self-monitoring in graduate medical education. Language conveys concepts, ideas, emotions, and information with the goal that the interpretation matches the intended meaning. Clearly defined terminology supports this goal. The scoping review and think aloud methods paper demonstrate the complexity of language and the confusion that may arise in absence of explicitly defined terminology.

The term “self-monitoring” only appeared in five of the included articles. The concept described, observed, and/or studied in each was self-monitoring; however, over 18 different terms were used to describe the concept and only seven included a clear definition. The lack of consistent nomenclature makes searching for research on self-monitoring quite challenging. A term such as “reflection-in-action” is specific enough that it could be included as part of a search involved Medical Subject Headings (MeSH). However, many of our included studies (n=32, 55%) used terms that lack specificity and, at worst, represent different concepts, such as “self-assessment”. Health professions education research, like all research, must build upon prior research, such that the community adds, often incrementally, to the collective knowledge. Heterogeneous terminology makes finding prior work and, therefore, building from it, very challenging.

When a new field or idea develops, of course, terminology is likely to be heterogeneous. Exactly when ideas should be consolidated into precise, consistent, clearly defined terminology is likely only answerable in hindsight. Nonetheless, a commitment of authors and researchers

to clearly defining the language chosen to represent ideas, such as self-monitoring, would go a long way in the evolution of universal understood terminology.

The think aloud protocol, on the other hand, suffers less from the heterogeneity of terminology to describe the same thing, as in self-monitoring, and more from the heterogeneity in meaning of the think aloud protocol. The term "think aloud" is inviting. Without reading more about the method, a researcher might feel like they understand what a think aloud is. On the one hand, this is great! The term is intuitive. On the other hand, the intuitiveness may lead to heterogeneity in application of the think aloud protocol and, as such, create definitional uncertainty. For example, Ericsson and Simon, co-creators and champions of the think aloud protocol, describe the inseparability of the think aloud interview and protocol analysis (Ericsson, 2002; Ericsson & Simon, 1980, 1981). However, the think aloud interview has been applied in various contexts using alternative analysis techniques without complement of protocol analysis (Croft et al., 2018; Gardin, n.d.; Pinnock et al., n.d.; Richardson et al., 2018). Some of these modifications have clearly occurred as part of the intentional and natural evolution of a research method (Göransson et al., 2007). On the other hand, the intentionality of other adjustments to the method is less clear (Richardson et al., 2018). In other words, the adjustments to the method, in some cases, may have been because of a lack of clear understanding of the think aloud protocol, creating unintentional variation that adds to the definition confusion. To clarify, methods should evolve, and variations are important for answering different research questions; however, researchers should conspicuously state deviations within their research methods and consider assigning new terminology to differentiate their method from the status quo. Referencing methods papers is an ideal way to

ensure high fidelity implementation of a particular research method. Unfortunately, methods papers are not always available. In the absence, health professions education researchers can optimize their research methods through replicating methods from high quality research papers with justifications for modifications as needed.

The think aloud protocol and the results from the scoping review synergize to provide a roadmap for future research on self-monitoring. The current literature on self-monitoring in GME focuses on post-task self-judgments (n=41, 71%) of performance on procedural skills, management, or interpretation. Yet, pre-task and during action self-judgment would seem most critical for patient safety during procedural skills and/or decisions around management and interpretation. A self-judgment of poor performance after the action will not help the patient under-the-knife or with their discharge paperwork in hand, even if it may improve future practice. Additionally, while the after action self-judgment of procedural skills suggests decent accuracy, findings within a procedural context do not uniformly translate to non-procedural contexts (Bonrath et al., 2015; Casswell et al., 2016; De Blacam et al., 2012; Ganni et al., 2017, 2018; Jamshidi et al., 2009; Mandel et al., 2005; Osborne et al., 2014; Quick et al., 2017; Scaffidi et al., 2018; Trajkovski et al., 2012; Veaudor et al., 2018; Vyasa et al., 2017; Ward et al., 2003). We need to unravel the thought process of self-monitoring to better understand how self-monitoring may translate to non-procedural tasks, to understand the practicality of pre- and during task self-monitoring, and to identify opportunities for improving calibration and frequency of self-monitoring in practice. In other words, we need to better *characterize* self-monitoring.

Slightly less than half (n=25, 43%) of the studies in our scoping review characterized self-monitoring or focused on the process of self-monitoring. And, of these, most focused (n=18, 31%) on factors that may influence self-monitoring or factors influenced by self-monitoring, rather than the thinking process of self-monitoring. Only seven studies (12%) tried to better understand the thinking processes of self-monitoring through interviews and focus groups. Only one (2%) used a think aloud interview (Surry et al., 2017). *Observing* the thinking process, through think aloud interviews, may well offer new insights into opportunities to calibrate and encourage self-monitoring.

For example, we are interested in performing a think aloud interview on GME trainees using clinical vignettes that require diagnostic reasoning and management with the goal of evaluating how thought process influences the accuracy of self-judgment of performance. In one planned experiment, we plan to compare the self-judgment of performance of each participant in a sleep replete and sleep deprived state. In a second, we plan to compare the process of self-monitoring, using a think aloud interview, between high and low performing trainees. With this combination, we hope to identify features of high- and low-quality self-monitoring, as an opportunity to develop interventions to improve self-monitoring.

These two examples are built around simulated clinical tasks. We suspect, as aligns with situated cognition theory, that the clinical context influences self-monitoring. And, therefore, we have also tentatively planned a study where participants would perform self-judgments of performance in the context of outpatient clinic encounters with selected opportunities for think aloud interviews during their self-judgments within the workplace environment. Aligning the simulated clinical tasks with real outpatient clinical tasks would offer a comparison for self-

monitoring within a simulated versus workplace environment. Moreover, the differences may offer additional opportunities for intervention. For example, imagine that self-monitoring is impaired in a noisy clinical space. A simple intervention, such as a purposeful bathroom break or walk down a quiet corridor in response to a trigger or cue, may reduce cognitive errors. This would be akin to what has previously been described as "slowing down" among surgeons in the operating room (Moulton et al., 2010). The think aloud protocol offers opportunities to expand the health profession education communities understanding of self-monitoring, laying further groundwork for interventions to improve the effectiveness of self-monitoring in the environment that matters most – the workplace.

Military Relevance

Military medical officers at every rank will face diagnostic or management uncertainty. Some military medical officers will experience this feeling in a large military treatment facility, like their civilian physician counterparts. Others will find themselves in remote and austere operational environments and similar levels of uncertainty. Ignoring or failing to recognize uncertainty may compromise patient safety and even lead to harm. Self-monitoring is a cognitive tool that may help to reduce patient harm. In this thesis, we have characterized the current understanding of self-monitoring and offered an underutilized method for furthering our understanding of self-monitoring. We have laid a foundation for future work to further understand the influences of effective self-monitoring so that interventions can be designed to optimize self-monitoring at the individual level.

Consider a general medical officer (GMO) stationed with the marines who encounters an 18-year-old recruit with nausea and vomiting found to have a sodium of 108 mmol/L. The GMO appropriately triages the patient as "sick" and requests transport to the nearest hospital. Prior to departure, the GMO starts 543 mL/hour of normal saline based on a mobile-device calculator for sodium correction rate. Fast forward 1-2 weeks, the 18-year-old recruit has devastating and irreversible neurologic injury from osmotic demyelination syndrome. The recruit arrived at the hospital 4 hours after first contact with the GMO with a serum sodium of 130 mmol/L. Perhaps coaching on self-monitoring during the GMO's GME training could have prevented this adverse outcome. The GMO may have looked at that number generated by the sodium correction calculator of 543 mL/hour and thought, "that seems high … let me pause and think about this … correcting sodium too fast can definitely be bad … let me ask for help". Prior to transport, the GMO may have reached out to the hospital and asked to speak with the admitting physician or a nephrologist. Together, they may have come up with a different plan. A different plan that could have change a marine recruit's life for the better.

Similar events could happen with hospital corpsmen, independent duty corpsmen, medical service corps officers and providers (e.g., physicians assistants), nursing corps, or seasoned subspecialists. Similar events could happen in any hospital in the country, including the military treatment facilities. None can know everything. Medicine is humbling. To reduce cognitive errors, we must recognize when we need to find help. The anecdote above, based on a true story, is extreme and displays the holy grail that self-monitoring could be. We are not there. Yet, investing in the research, such as this thesis, to move us closer is a small price to pay for the potential reward.

Conclusions

Self-monitoring is a complex yet appealing phenomenon that seems promising as a research focus to improve the quality of GME and the safety of clinical practice. The current literature is riddled with challenges, most notably a lack of clear definitions and terminology. The think aloud protocol is an underutilized method for researching thought processes, such as self-monitoring, and may well aid in addressing some of the gaps in the current self-monitoring literature. Like the term self-monitoring, the term think aloud interview has heterogeneity in implementation. Through this thesis, we hope that we have clarified the current landscape of self-monitoring in graduate medical education, described a feasible best-practice for implementing the think aloud protocol, and suggested future directions for self-monitoring research in health professions education that leverage the synergy between the gaps in the literature and the utility of the think aloud protocol.

Appendix 1: Search strategy for scoping review (Chapter 2)

TOTAL Results: 6,437 - 2309 duplicates = 4128 search results
Ovid Medline Search Strategy, 04/16/2019 - 2,231 results

Self Directed Learning as Topic/ OR ((self-direct* or self-regulat*) and learning).ti,ab.
(self-assessment* and self-monitoring).mp.
((exp *Awareness/ or exp *Metacognition/ OR exp *Self Assessment/) and (exp *Clinical Competence/ or exp Learning/ OR exp Teaching/))
((awareness or metacognition or self-assessment or self-monitoring) adj10 (learning or elearning or clinical competenc* or interactive-feedback or interactive-feed-back or medical competenc* or medical-knowledge or surgical competenc* or teaching)).mp.
or/1-4
(exp *Clinical Clerkship/ or exp *Clinical Competence/ or exp *Education, Medical/ or exp *Faculty, Medical/ or exp *Physicians/ or exp *Students, Medical/ or (clinician$ or clinical-clerkship$ or doctor$ or faculty or medical student* or physician$ or practitioner$ or resident$ or surgeon$).ti. Or (clinical-competenc* or clinical-environment or postgraduate or post-graduate).ti,ab.)
5 and 6
Limit 7 to english language
Embase Search Strategy, 04/16/2019 – 2297 results
('self-directed learning readiness scale'/exp OR 'self-directed learning'/exp OR (('self direct*':ti,ab,kw OR 'self regulat*':ti,ab,kw) AND learning:ti,ab,kw))
self-assessment* AND self monitoring
('awareness'/exp/mj OR 'metacognition'/exp/mj) AND ('clinical competence'/exp/mj OR 'learning'/exp/mj OR 'teaching'/exp/mj)
(awareness OR metacognition OR self-assessment OR self-monitoring) NEXT/10 (learning OR elearning OR clinical-competenc* OR interactive-feedback OR interactive-feed-back OR medical-competenc* OR medical-knowledge OR surgical-competenc* OR teaching)
#1 OR #2 OR #3 OR #4
'clinical competence'/exp/mj OR 'medical education'/exp/mj OR 'physician'/exp/mj OR 'medical student'/exp
(clinician* OR clinical-clerkship* OR doctor OR doctors OR faculty OR medical-student* OR physician* OR practitioner*OR resident OR residents OR surgeon*):ti,kw
(clinical-competenc* OR clinical-environment OR postgraduate OR post-graduate):ti,ab,kw
#6 OR #7 OR #8
#5 AND #9

Eric.ed.gov Search Strategy, 04/16/2019
(Note:Limited to Peer reviewed only)
“Self directed learning” AND (medicine OR medical) – 89 results
“Self monitoring” AND (medicine OR medical) – 230

"self regulated learning" AND (medicine OR medical) – 24
"self regulation learning" AND (medicine OR medical) – 2

Updated PsycINFO Search Strategy, 04/16/2019, 797 records
Exp self-regulated learning/ OR ((self-direct* or self-regulat*) and learning).ti,ab.
(self-assessment* and self-monitoring).mp.
((exp *Awareness/ or exp *Metacognition/ OR exp *Self-Evaluation/) and (exp Learning/ OR exp Teaching/))
((awareness or metacognition or self-assessment or self-monitoring) adj10 (learning or elearning or clinical competenc* or interactive-feedback or interactive-feed-back or medical competenc* or medical-knowledge or surgical competenc* or teaching)).mp.
or/1-4
(exp *Medical Education/ or exp *Physicians/ or exp *Medical Students/ or (clinician$ or clinical-clerkship$ or doctor$ or faculty or medical student* or physician$ or practitioner$ or resident$ or surgeon$).ti. or (clinical-competenc* or clinical-environment or postgraduate or post-graduate).ti,ab.)
5 and 6
Limit 7 to english language

Web of Science Search Strategy, 04/19/2019 - 767 results

((self-direct* or self-regulat*) and learning) OR (self-assessment* and self-monitoring)
(awareness or metacognition or self-assessment or self-monitoring) NEAR/10 (learning or elearning or clinical-competenc* or interactive-feedback or interactive-feed-back or medical-competenc* or medical-knowledge or surgical-competenc* or teaching)
#2 OR #1
TI=(clinician* or clinical-clerkship* or doctor* or faculty or medical-student* or physician* or practitioner* or resident* or surgeon*) OR TS=(clinical-competenc* or clinical-environment or postgraduate or post-graduate)
#4 AND #3 AND
#4 AND #3 (NOTE: Refine by English language)

Appendix 2: Coding sheet for scoping review (Chapter 2)

· Primary Coder: Investigator completing initial coding. Any cells of uncertainty, requesting more thorough second review, should be highlighted in **YELLOW**
· Secondary Coder: Investigator assigned coding verification. Any discrepancies should get highlighted (i.e. highlight cell in **RED**) for discussion and consensus.
· First Author: Full name, as listed, of the paper's first author.
· Year: Year of publication.
· Title: Article Title.
· Journal or Dissertation.
· Journal Name: Name of the journal (if applicable) in which article is published.
· Methodology:
    - Quantitative: Methods and results focus on numerical collection and analysis
    - Qualitative: Methods and results focus on verbal or written collection w/o any comparative statistical analyses.
    - Mixed Methods: Combines the two approaches above.
· Comparison Group: A group of participants without an intervention or with a different intervention.
· Randomized: Participants randomized into different study arms.
· Study Population:
    - Single: All the same year of trainees.
    - Mixed: Different years of trainees.
· Study Population: List all the training years and any other populations included in the study.
· Average years in training for GME participants: Record if reported. If not reported and proportion of different training years reported, calculate the average years of training using training year (e.g. 10 PGY1 and 10 PGY2 average training year (10*1+10*2)/20=1.5)
· Number of study participants: Total number of all study participants.
· Age of participants: Average (mean) age of study participants w/ standard deviation, if presented)
· Percent of GME trainees: The percentage of participants that are GME trainees.
· Percent Female GME Trainees: The percentage of GME trainees that are female.
· Medical Specialty or Specialties: List GME training program types that are represented (e.g. internal medicine, pediatrics, psychiatry, general surgery, etc.)
· Single or Multi-Institutional
    - Single: Participants all from one GME program
    - Multi-institutional: Participants from 2 or more GME programs
· Study Geographic Location:
    - USA: GME program in USA
    - Other Country: GME program outside USA
    - Multiple: GME programs in multiple countries.

· State(s) or Country(ies): List the states or countries that the GME program is located in (Note: If GME program located in Maryland, but has hospital affiliations in other states, still list as Maryland)
· Rural, Urban, Mixed:
    - Rural: Hospital in population area of less than 50,000
    - Urban: Hospital in population area of greater than or equal to 50,000
    - Mixed: Hospitals in population areas that meet both definitions above.
    - Unknown: Site of research not listed or not clear based on author affiliations.
· Explicit Accreditation, Certifying, or External Guiding Standard: Yes or No (Note: These are organizations or large bodies that published recommendations or standards for implementation in education or clinical practice)
· What is/are the guiding standard(s): List the organizations.
· What year(s) is/are the guiding standard(s) from: List the years of the guiding standards.
· Was there another external motivation for the authors' study? Free text any central motivating factor to authors creation of the study, typically from the background section.
· Study Setting:
    - Inpatient: Education/practice in setting of hospitalized patients outside operating room
    - Outpatient: Education/practice in setting of non-hospitalized, ambulatory care, patients, also outside the operating room
    - Operating Room: Education/practice predominantly contained to the operating room.
    - Simulation: Education/practice in simulation center/setting.
    - Other: Not one of the above.
    - NOTE: If mixed setting, then okay to list multiple.
· Care Type:
    - Primary: Internal Medicine, Family Medicine, General Practice, Pediatrics, Psychiatry, and General, Outpatient Ob/Gyn, Emergency Medicine
    - Procedural/Surgical: Patient interactions or practiced skills for use within operating room environment
    - Consultant/Specialist: Neither of the above categories
· Task Type:
    - Procedural: Execution of motor-based skill.
    - Interpretation: Decision(s) about meaning of data.
    - Management: Decision(s) about therapeutic interventions or diagnostic testing.
    - Team Care Delivery: Collaboration with team to deliver care.
    - Communication: Patient/Family-provider or one-on-one provider-provider interaction
    - Other: None of the above.
· Explicitly stated study purpose(s) or question(s): Yes or No.

· Study Purpose or Question: Direct quote preferred. If purpose clear, but not explicitly stated, please state in this block. If direct quote, use quotations.
· Explicitly stated learning theory framework: Yes or No.
· If yes, what learning theory or framework was used: State the learning theory and/or most appropriate learning theory bucket.
· Learning Theory Citation: Copy and pasted the citation for the learning theory used.
· Terminology Used as Self-Monitoring Equivalent: Use verbatim terms from article. Okay to list multiple.
· Definition (if verbatim, use quotes)
· Position of self-monitoring in study
    - Pre-task: Precedes task.
    - During task: Coincides with task.
    - Post-task: Occurs after completion of task
· Role of Self-Monitoring
    - Characterized: Focuses on the process of self-monitoring or development of successful/accurate self-monitoring
    - Encouraged: Focuses on increasing frequency of self-monitoring.
    - Taught: Focuses on increasing rigor/impact of self-monitoring.
    - Measured: Quantifies accuracy or impact of self-monitoring.
· Self-monitoring is concurrent, retrospective (e.g. video review, self-audit), or both:
    - Concurrent: Occurs in the same time space as action being self-monitored
    - Retrospective: Not temporally connected to action being self-monitored
    - Both: Includes both of the above
· How was self-monitoring characterized/encouraged/taught/measured: Free text the role of self-monitoring within the study.
· Influence of Self-Monitoring:
    - Self-Observation: Recognizing one's own thoughts, attitudes, or behaviors
    - Self-Judgment: Grading accuracy/appropriateness of one's own thoughts, attitudes, or behaviors.
    - Self-Reaction: Reinforcing, planning, or implementing change in thoughts, attitudes, or behaviors.
· Primary Study Outcomes/Results: Briefly describe primary study outcomes or results and then focus on outcomes related to self-monitoring.
· Study Limitations: Briefly describe any limitations that the authors mention and any additional limitations that you see to the study.
· Suggested best practices: Briefly describe how the authors suspect their findings could influence practice or future research.
· Statistical or Functional Impact: Measurements or impacts of self-monitoring qualitative or quantitative.
· Was the impact of self-monitoring:
    - Positive: Self-monitoring outcomes from the study suggest that self-monitoring is or could be useful.

- Neutral: Self-monitoring outcomes from study do not that self-monitoring may not be useful.
- Negative: Self-monitoring outcomes from study suggest that self-monitoring is not or would not be useful.

- Quality Rubric (Grade on different sheet and will autopopulate)
  - Quantitative Studies or Mixed Method Studies (MERSQI)
    - § Study Design
      - Single-group cross-sectional or posttest only (1)
      - Single group pretest and posttest (1.5) Also include series or repeated measures.
      - Non-randomized (2)
      - Randomized (3)
    - § Sampling Institutions
      - Single (0.5) GME program
      - Two (1) GME programs
      - Three or More (2) GME programs
    - § Sampling Response Rate
      - N/A: Only if response rate truly does not apply.
      - Less than 50% (0.5) or not reported (in study where it should have been reported) Of eligible participated or were analyzed
      - 50-74% (1) Of eligible participated or were analyzed
      - 75-100% (1.5) Of eligible participated or were analyzed
    - § Type of Data: Averaged (mean) where multiple types of data collected for outcome.
      - Assessment by study participant (1): Self-observation in this bucket, even if during video review. If data collected from interview, then (2).
      - Objective, including observation (3)
    - § Data Analysis, Sophistication
      - Descriptive Analysis only (1): Frequency, mean, median, or mode and ranges/variance only.
      - Beyond Descriptive Analysis (2): Any statistical tests for comparison.
    - § Data Analysis, Appropriate
      - Yes (1): Reasonable
      - No (0): Statistical error or erroneous missing data
    - § Outcome: Where multiple true for outcome, average taken.
      - Satisfaction/Attitude/Perception/Opinion (1): Self-reporting given heterogeneity of data, in all cases considered 1 and averaged with type of self-rating.
      - Knowledge /Skill (1.5): Knowledge or skills in non-situated environment (i.e. element of inauthenticity; includes simulation)
      - Behavior (2): Workplace, situated knowledge, skill, or attitude

- Patient/Healthcare Outcomes (3): Impact beyond participants to patients and their health

§ Validity Evidence for Instrument: Total number of points below. If multiple instruments used, calculate score for each instrument and record the average (mean). Additionally if same tool used for multiple populations, then separate validity scores for each population averaged.

- N/A: No instrument used.
- Content (1 point): Evidence presented/referenced for information included in the instrument (e.g. literature, expert review/consensus)
- Internal Structure (1 point): Evidence presented/referenced for reliability, reproducibility, test-retest, internal consistency
- Relationship to other variables (1 point): Evidence presented/referenced for instrument correlating with related variable or not correlating with unrelated variables.

○ O'Brien SRQR: Qualitative Studies or Mixed Methods Studies (For each of the following categories assign 0 points for Not included/met, 0.5 for partially included/met, and 1 point for completely included/met). If a mixed methods study, focus on the qualitative portion of the study.

§ Title: Concise description of the nature and topic of the study Identifying the study as qualitative or indicating the approach (e.g., ethnography, grounded theory) or data collection methods (e.g., interview, focus group) is recommended

§ Abstract: Summary of key elements of the study using the abstract format of the intended publication; typically includes background, purpose, methods, results, and conclusions

§ Introduction

- Problem Formulation: Description and significance of the problem/phenomenon studied; review of relevant theory and empirical work; problem statement
- Purpose or Research Question: Purpose of the study and specific objectives or questions

§ Methods

- Qualitative approach and research paradigm: Qualitative approach (e.g., ethnography, grounded theory, case study, phenomenology, narrative research) and guiding theory if appropriate; identifying the research paradigm (e.g., postpositivist, constructivist/ interpretivist) is also recommended; rationale
- Research characteristics and reflexivity: Researchers' characteristics that may influence the research, including personal attributes, qualifications/experience, relationship with participants, assumptions, and/or presuppositions; potential or

actual interaction between researchers' characteristics and the research questions, approach, methods, results, and/or transferability

- Context: Setting/site and salient contextual factors; rationale
- Sampling Strategy: How and why research participants, documents, or events were selected; criteria for deciding when no further sampling was necessary (e.g., sampling saturation); rationale
- Ethical issues pertaining to human subjects: Documentation of approval by an appropriate ethics review board and participant consent, or explanation for lack thereof; other confidentiality and data security issues
- Data collection methods: Types of data collected; details of data collection procedures including (as appropriate) start and stop dates of data collection and analysis, iterative process, triangulation of sources/methods, and modification of procedures in response to evolving study findings; rationale
- Data collection instruments and technologies: Description of instruments (e.g., interview guides, questionnaires) and devices (e.g., audio recorders) used for data collection; if/how the instrument(s) changed over the course of the study
- Units of study: Number and relevant characteristics of participants, documents, or events included in the study; level of participation (could be reported in results)
- Data processing: Methods for processing data prior to and during analysis, including transcription, data entry, data management and security, verification of data integrity, data coding, and anonymization/deidentification of excerpts
- Data analysis: Process by which inferences, themes, etc., were identified and developed, including the researchers involved in data analysis; usually references a specific paradigm or approach; rationale
- Techniques to enhance trustworthiness: Techniques to enhance trustworthiness and credibility of data analysis (e.g., member checking, audit trail, triangulation); rationale

§ Results

- Synthesis and interpretation: Main findings (e.g., interpretations, inferences, and themes); might include development of a theory or model, or integration with prior research or theory
- Links to empirical data: Evidence (e.g., quotes, field notes, text excerpts, photographs) to substantiate analytic findings

§ Discussion

- Integration with prior work, implications, transferability, and contribution(s) to the field: Short summary of main findings; explanation of how findings and conclusions connect to, support, elaborate on, or challenge conclusions of earlier scholarship; discussion of scope of application/ generalizability; identification of unique contribution(s) to scholarship in a discipline or field
- Limitations: Trustworthiness and limitations of findings (Table

§ Other

- Conflicts of interest: Potential sources of influence or perceived influence on study conduct and conclusions; how these were managed
- Funding: Sources of funding and other support; role of funders in data collection, interpretation, and reporting

- Other Coder Comments: Free text, any comments.
- Dual Coding Complete & Reconciled: Yes or No.

Appendix 3: Detailed coding of included studies in scoping review (Chapter 2)

Author (Year)	Journal Name	Study Design (Methodology / Comparison Group / Randomized)	Number of study participants	Percent GME Trainees (Residents or Fellows)	Study Purpose or Question (if verbatim, use quotes)	Study Quality Rating (MERSQI/SRQR)	Learning theory or Conceptual Framework Used	Terminology Used as "Self-Monitoring" Equivalent	Task Type	Timing of Self-Monitoring	Type of Self-Monitoring	Role of Self-Monitoring in Study (characterized, encouraged, taught, or measured)	How was self-monitoring used?	Brief summary of results relevant to self-monitoring
Bob Wu (2009)	Journal of Surgical Education	Quantitative / No / No	7	100.00%	"The authors now present a rapid, web-based, resident self-driven, and quantitative outcome-assessment tool for the objective documentation of PBLI [practice based learning and improvement], as well as a tool for discovering resident-specific deficiencies to stimulate targeted self-study."	42%/ -	None	Self-Drive, Self-Assessment, Self-reporting	Interpretation, Management	Concurrent post-task	Self-Observation	Encouraged, Measured	Encouraged: Case-log portal requiring initial diagnosis and treament Measured: Follow-up diagnosis and treatment with self-rated accuracy	For more novice learners, diagnostic and management accuracies improved each quarter
Jan Wouda (2014)	Patient Education and Counseling	Quantitative / No / No	44	100.00%	"We investigated the effects of an innovative program for communication self-assessment supplemented with supervisor feedback, on residents' communication-competency awareness and on their communication competency in general and their patient-education competency in particular."	73% / -	None	Self-Assessment, Self-Awareness	Communication	Retrospective post-task	Self-Judgement, Self-Reaction	Taught, Measured	Taught: Feedback and dialogue between trainees and faculty to improve self-monitoring Measured: Self-rating with instrument compared to expert rating	Combination of self-monitoring and expert feedback improved trainee communication skills and increased the number of learning objectives as targets for onging self-monitoring
Parth Vyasa (2016)	Journal of Surgical Education	Quantitative / Yes / Yes	30	100.00%	"To achieve these aims, we designed a study with the following 3 overarching goals: (1) to examine the accuracy of residents' assessments of their endoscopic skills, (2) to investigate if accuracy improves over time and practice, and (3) to compare the efficacy of 3 interventions—practice only (PO), SO, or expert observation (EO)—on self-assessment accuracy."	70% / -	None	Self-assessment	Procedural	Concurrent and retrospective post-task	Self-Observation, Self-Judgement	Measured, Taught	Taught: Self-rating calibrated with expert rating or video (self-)observation Measured: Self-rating using instrument compared with expert rating	The accuracy of self-judgment significantly improved with subsequent endoscopy performance with expert observation and feedback, improved (without significance) with self-directed video observation, and did not change in the control group

Martin Veaudor (2018)	BMC Medical Education	Quantitative / Yes / No	34	67.60%	"we sought to determine whether a high-fidelity FB simulation self-training program would allow novice bronchoscopists to acquirecompetencies similar to those of trained bronchoscopistsas concerns the visualization of the bronchial tree and the identification of its anatomical elements."	77% / -	None	"Felt Ready"	Procedural	Concurrent during task and post-task	Self-Judgement	Characterized	Characterized: Participants in one of three groups were allowed to practice until they felt ready for their final evaluation. The other groups followed a strict protocol.	Participants allowed to practice allows until they self-judged themselves ready had significantly more improvements on their final evaluation compared to the other groups, even outperforming more participants with more experience in some categories. They also had highest satisfaction among the groups.
Valerie van den Eertwegh (2014)	PLOS One	Qualitative / No / No	12	100.00%	“How can the learning process that GP residents undergo in order to become skilled communicators be described?”	- / 79%	Reflective-Impulsive Model of Social Behavior	Reflecting, Conscious	Communication	Retrospective post-task	Self-observation, Self-Reaction	Characterized	Characterized: Post-task interview with simulated recall techniques to understand how trainees decided to communicate.	The following themes were identified as influencers of communication decisions: (1) Confrontation w/ the effect of a behavior, (2) Becoming conscious of own behavior, (3) Searching & Receiving Alternative Behavior, (4) Personalization, (5) Internalization and integration, (6) Two overall conditions influencing learning process -- supervisor relationship & cognitive load
Tomce Trajkovski (2011)	Canadian Journal of Surgery	Quantitative / No / No	34	50.00%	"The purpose of the present study was to examine orthopedic residents’ self-assessment of surgical involvement and competence inperforming primary total knee (TKA) and total hip arthroplasty (THA)."	76% / -	None	Self-Assessment	Procedural	Concurrent post-task	Self-observation, Self-judgement	Encouraged, Measured	Encouraged: Instrument to promote documentation of case involvement, complexity, and competency Measured: Self-rating compared with expert rating	Strong correlation between expert and trainee ratings for involvement and competency with greater rates of underestimation.
Luke T Surry (2017)	Medical Education	Qualitative / No / No	8	62.50%	"To evaluate whether MCQ have expected elements of real-world clinical reasoning."	- / 93%	Dual-Processing Theory	None	Interpretation, Management	Concurrent post-task	Self-observation	Characterized	Characterized: Participants articulated thinking processes, which included self-monitoring.	Trainees are capable of articulating their thinking processes, which include self-monitoring, immediately after answering a multiple choice question.

Paula Ravitz (2013)	Academic Psychiatry	Quantitative / Yes / No	26	100.00%	"The objectives of the current study were to determine 1) whether family medicine trainees' competence in thera- peutic communication improves over the course of the CCSE intervention; 2) whether the trainees' self-efficacy for therapeutic communication improves over the course of the intervention; 3) whether this improvement differs from changes resulting from residency training-as-usual; and 4) whether this improvement is maintained up to 6 months later."	74% / -	Experiential Learning Theory, Adult Learning Theory	Self-assessment, Self-Reflection, Self-observation	Communication	Concurrent and retrospective post-task	Self-Observation, Self-Reaction, Self-Judgement	Taught, Measured	Taught: 1-month intervention with iterative task performance, self-rating, feedback, and coaching. Measured: Self-rating of confidence compared with expert rating of competence	Communication self-efficacy (i.e., confidence in his or her ability) and expert rated competence improved after intervention with iterative practice, self-monitoring, and coaching.
Jacob A Quick (2017)	Journal of Surgical Research	Quantitative / No / No	14	100.00%	"We aimed to determine whether differences exist between resident self-assessment and observer evaluation of technical skill when performing a laparoscopic cholecystectomy."	79% / -	None	Self-Evaluation	Procedural	Concurrent post-task	Self-Judgement, Self-Observation	Measured	Measured: Self-rating compared with expert rating	Strength of correlation increased with number of years of training. Strong correlation between faculty and PGY5 trainees. Across all groups, cases at the extremes of difficulty (easiest, hardest) had strongest correlation between trainees and faculty.
Elisabeth AM Pelgrim (2012)	BMC Medical Education	Quantitative / No / No	54	100.00%	"We investigated the frequency of different types of comments invited in the form: self-reflection by the trainee, feedback from the trainer and an action plan proposed by both trainer and trainee."	67% / -	Social Self-Regulated Learning Theory	Self-reflection	Communication, Management, Interpretation	Concurrent post-task	Self-Reaction	Encouraged	Encouraged: Trainee self-reflection section on feedback tool used by faculty to evaluate competence in observed clinical encounters	About half of the trainees used the self-reflection space and when used the comments tended to be specific. The pairing of the faculty-trainee may influence rate of self-reflection.

Bridget R. O'Connell-Long (2016)	Journal of Surgical Research	Quantitative / No / No	44	100.00%	"The aim of this study was to investigate whether surgical residents had successfully mastered bladder catheterization"	68% / -	None	Self-confidence	Procedural	Concurrent pre-task and post-task	Self-Judgement, Self-Observation, Self-Reaction	Measured, Characterized	Characterized: Comparison of pre-task and post-task self-monitoring. Measured: Comparison of number of errors committed during task and self-ratings of confidence.	Pre-scenario ratings of confidence were higher than post-scenario and had better (albeit weak) correlation with number of errors committed. Higher ratings correlated with more rapid requests for help/assistance.
Krishna Moorthy (2006)	The American Journal of Surgery	Quantitative / No / No	27	100.00%	"The aim of this study was to compare the assessment by independent observers with the self-assessment of performance by trainees for technical and nontechnical/team skills during a simulated procedure in a simulated operating theater (SOT)."	67% / -	None	Self-Assessment	Procedural, Team Based Care	Concurrent post-task	Self-observation, Self-judgement	Measured, Characterized	Characterized:Influence of training experience on self-judgment accuracy. Measured: Self-ratings compared to expert ratings	Strong correlation between expert and trainee ratings for technical skills with improved correclations in trainees with more experience. Poor to moderate correlation of self-rating of non-technical skills with greater correlation with less experienced trainees.
Lynn S. Mandel (2005)	American Journal of Obstetrics and Gynecology	Quantitative / No / No	74	100.00%	"The purpose of this study was to examine resident assessments of their own proficiency on a variety of surgical bench procedures and to compare their ratings with the ratings of trained faculty observers who used instruments that have been shown to be reliable and valid."	86% / -	None	Self-Assessment	Procedural	Concurrent post-task	Self-Observation, Self-Judgement	Measured	Measured: Self-rating compared with expert rating	Strong correlation between faculty and trainees with trainees tending to underestimate.
Sirirat Ularntinon (2016)	Academic Psychiatry	Quantitative / No / No	1	100.00%	"This brief article describes the SELF and explains its usefulness in residents' professional development."	43% / -	Bennett-Levy Clinical Learning (Cognitive Learning Theory)	Self-Reflection	Communication, Management	Concurrent post-task	Self-Observation, Self-Judgement, Self-Reaction	Taught, Measured	Taught: Faculty coaching and structured instrument Measured: Self-rating compared with expert rating	Combination of self-rating and coaching helped to create learning objectives and roadmap.

Jennifer L. Plant (2012)	Advances in Health Sciences Education	Mixed / No / No	16	100.00%	"we examined the process of informed self-assessment in action in a specific educational context, with a goal to better understand how, why, and to what extent resident physicians adjust their self-assessment based on external information."	59% / 71%	Sociocultural learning theory	Self-Assessment	Team Based Care, Management	Concurrent post-task and retrospective during and post-task	Self-Observation, Self-Judgement, Self-Reaction	Characterized, Measured	Characterized: Interviewed to understand process of self-monitoring. Comparison of concurrent and retrospective self-monitoring. Measured: Self-rating compared with expert rating.	Trainees tended to underestimate their performance. Retrospective ratings were generally lower than concurrent ratings. No correlations reported.
Alan J. Osborne (2014)	Medical Teacher	Mixed / No / No	25	100.00%	We aimed to introduce the PBA as a self-assessment tool and to evaluate the learning outcomes in order to answer the following questions: ''Is self-assessment PBA a valid tool when compared to an external assessment PBA? Do self-assessment PBAs identify the same learning needs and do they identify change in clinical practice?''	70% / 48%	Experimental Learning Theory	Self-Assessment	Procedural	Concurrent post-task	Self-Observation, Self-Judgement, and Self-Reaction	Measured	Measured: Self-rating compared with expert rating	Self-rating correlated more strongly with experts based on years of training. Experts and self-identified action plans were similar in items, but trainees included more non-technical objective (v. technical).
Karen D. Konings (2016)	Academic Medicine	Quantitative / Yes / Yes	64	100.00%	"We investigated whether use of the reflection app and participation in the coaching sessions increased intensity and frequency of reflection. We also tested whether these two types of support together fostered additional reflection."	76% / -	None	Reflection, Self-Monitoring, Self-recording	Communication, other (professionalism)	Concurrent during and post-task, retrospective post-task	Self-Observation, Self-Reaction	Characterized, Encouraged, Taught, Measured	Characterized: Comparison of rates of documented self-monitoring with different interventions (e.g., coaching, mobile app, reflective writing) Encouraged: All verbally prompted to reflect on learning moments. Some also had augmentation with mobile application. Taught: Some trainees received coaching. Measured: Frequency of self-monitoring	Mobile applications increased frequency of self-monitoring. Coaching improved content of self-monitoring.

Arjun D. Koch (2011)	American Journal of Gastroenterology	Quantitative / No / No	19	100.00%	N/A	50% / -	Osborn–Parnes Creative Problem Solving Process	Self-assessment, Self-evaluation	Procedural	Concurrent post-task	Self-observation, Self-judgement	Encouraged, Measured	Encouraged: Instrument to promote documentation of self-monitoring Measured: Self-ratings compared overtime and correlated with clinical documentation	Self-rating consistent with clinical documentation. Trainees had a high completion rate (91%) of self-monitoring instruments with low completion rate of learning plans (36%).
Young-Mi Kim (2002)	International Journal for Quality in Health Care	Mixed / Yes / Yes	60	100.00%	"The specific objectives are: (1) to determine if supervision and self-assessment help doctors to apply neuwly learned communication skills on the job and to improve those skills over time; and (2) to identify which activities (including supervision visits, audiotaped consultations, self-assessment, homework logs, and job aids) are effective and acceptable to doctors."	91% / 48%	None	Self-Assessment	Communication	Retrospective during task	Self-Observation, Self-Judgement, Self-Reaction	Encouraged, Measured, Taught	Encouraged: Instruments to promote documentation of self-monitoring Taught: Some participants received coaching Measured: Frequency of self-monitoring	All trainees communication skills improved overtime and the rate of improvement was greater in the group that received coaching. The freuqnecy of self-monitoring also seemed to influence rate of growth.
Anju Kapoor (2017)	Indian Pediatrics	Mixed / Yes / No	4	100.00%	The purpose of the present study was to assess the educational effectiveness of SNAPPS (Summarize history and findings; Narrow differentials; Analyze differentials; Probe preceptor about uncertainties; Plan management; Select case-related issues for self-study) model in comparison to traditional method for training pediatric residents in the OPD (outpatient department)-setting at a teaching hospital.	64% / 55%	Learner-centered model for presenting	Self-selected, Self-Directed	Management, Interpretation	Concurrent post-task	Self-Reaction	Encouraged, Taught	Encouraged: Implementation of new presentation style Taught: Instruction on new presentation style encompassing self-monitoring	Self-reflection occurred in 60% of encounters after intervention as compared to 0% prior to the intervention
Yinin Hu (2013)	Journal of Surgical Education	Quantitative / No / No	23	30.40%	"The purpose of this study is to utilize video self-assessment to compare perceived vs actual competency in the basic surgical techniques of suturing and knot tying."	62% / -	Social Cognitive Learning Theory	Self-Assessment, Self-Evaluation	Procedural	Retrospective during task	Self-Judgement	Measured	Measured: Self-rating compared with expert rating	Trainees found self-ratings and video observation valuable. Self-ratings did not correlate with faculty ratings and consistently trainees consistently overestimated.

Eric S. Holmboe (2005)	Academic Medicine	Quantitative / Yes / No	26	100.00%	"we sought to investigate the effects of a multifaceted experiential self-directed curriculum on quality improvement that involved residents' self-audit and reflection on their practice performances."	97% / -	None	Self-audit, Self-reflection	Management	Retrospective post-task	Self-judgement, Self-Reaction	Taught, Measured	Taught: Intervention group received instruction in self-audits and reflections using instrument. Measured: Comparison of patient outcomes between intervention and control groups	Patients assigned to participants in the intervention group delivered higher rates of comprehensive diabetic care.
Brian Hodges (2001)	Academic Medicine	Quantitative / No / No	24	100.00%	"We set out to determine whether the same problems of self-assessment so consistently replicated in the studies of Kruger and Dunning [i.e. poor calibration of self-assessment in least competent learners] could be demonstrated in family medicine residents."	68% / -	None	self-assessment	Communication, Management	Concurrent and retrospective post-task	Self-judgement	Measured	Measured: Self-rating compared with expert rating	Raw scores of self-rating had poor correlation; however, once the scores were normalized to individual mean, correlation was stronger
Christopher Hildebrand (2009)	Journal of General Internal Medicine	Mixed / No / No	80	100.00%	In our 5-year retrospective study, we assessed residents' improvement over time in their performance monitoring skills and outcomes, their skills at self-reflection, and their satisfaction with our PBLI practices. We also considered contextual effects of resident sex and clinic, and the feasibility and value of PBLI to the residency program as we assessed the impact of these practices on residents' commitment to improved patient care outcomes.	59% / 57%	None	Self-reflection	Management	Retrospective post-task	Self-Observation, Self-Judgment, Self-Reaction	Encouraged, Measured	Taught: Instructed in process of self-audit with instrument Measured: Frequency of practice of accepted preventative and chronic health measures.	Frequency of compliance increased with years of training and practice setting (higher at VA than university clinic).

Gabriel E. Herrera-Almario (2016)	The American Journal of Surgery	Quantitative / No / No	9	100.00%	"The purpose of this study is to determine the effect of video review on resident and attending assessments of a resident's laparoscopic surgical performance."	61% / -	None	Self-Assessment, Self-Monitoring	Procedural	Concurrent and retrospective post-task	Self-Observation, Self-judgement	Measured	Measured: Self-rating compared with expert rating. Concurrent and retrospective self-ratings and expert ratings compared.	Trainees concurrent and retrospective self-ratings were significantly different from concurrent and retrospective expert ratings respectively. Trainees underestimated their skills. Concurrent and retrospective self-ratings were not signitificantly different for trainees. Concurrent and retrospective self-rating were significantly different for experts.
Sarah Blissett (2018)	Advances in Health Sciences Education	Quantitative / No / No	22	100.00%	"The purpose of this study is to study the utilization of mental efort as a cue, the cue diagnosticity of mental efort, and the monitoring accuracy of certainty as a monitoring judgement using ECG interpretation in internal medicine residents as a model."	56% / -	Self-regulated learning theory, Cue Utilization Theory, Cognitive Load Theory	Monitoring judgements, Uncertainty/Certainty, confidence	Interpretation	Concurrent post-task	Self-judgement	Measured, Characterized	Characterized: Influence of mental effort on self-monitoring Measured: Self-rated certainty compared with accuracy of interpretation	Self-rated certainty had significant positive correlation with moderate effect size on the accuracy of interpretation. Mental effort had negative w/ moderate effect size (Beta + 0.343 w/ p<0.001). Mental effort had a negative correlation on self-rated certainty.
William Bond (2004)	Academic Medicine	Qualitative / No / No	15	100.00%	To assess the feasibility of learning cognitive error recognition during simulation training.	- / 55%	None	Metacognition, Cognitive Forcing	Team Based Care, Management	Concurrent post-task	Self-Observation, Self-Judgement	Characterized, Taught	Characterized: Factors influencing self-monitoring Taught: Participants received cognitive forcing training	Cogntive forcing training and mistakes seem to encourage self-monitoring.
Sylvain Boet (2011)	Critical Care Medicine	Quantitative / Yes / Yes	55	100.00%	"The purpose of our study was to test the relative effectiveness of a self-debriefing compared to instructor debriefing for learning nontechnical skills."	80% / -	None	Self-Assessment	Team Based Care	Retrospective during task	Self-Observation, Self-Judgement, Self-Reaction	Encouraged	Encouraged: Some participants had access to video review and rating instrument to guide self-monitoring. Others had experts guiding self-monitoring with probing questions.	Both groups had improved performance with interventions. Prior study with control group had no improvement. No difference in improvement based on intervention.

Richard Bounds (2013)	Academic Emergency Medicine	Quantitative / No / No	72	100.00%	"The objective of this study was to investigate the contributions of self-assessments and external feedback and how the two interact, in the for-mation of learning goals, as well as the reported follow-through on those goals, for performance improvement."	71% / -	None	Self-assessment	Interpretation	Concurrent post-task	Self-judgement, Self-Reaction	Measured	Measured: Self-rating compared with expert rating. Frequency of learning goals and implementation	Expert and self-ratings had significant, albeit weak, correlation. Self-monitoring had greater influence on learning goals than expert feedback. Learning goal implementation most likely when self-monitoring and feedback aligned.
Ryan Brydges (2016)	Journal of Graduate Medical Education	Qualitative / No / No	20	100.00%	"We studied how postgraduate year (PGY) 1 residents think about and regulate their learning, particularly whether and how they exhibit strategic mindfulness during training on 2 clinical scenarios (ie, easy and difficult) using a part-task LP simulator."	- / 86%	Self-regulated learning theory	self-monitoring, self-assessment	Procedural	Retrospective post-task	Self-observation, Self-Judgement	Characterized	Characterized: Interviewed to understand process of self-monitoring.	Self-monitoring seems to occur predominantly pre- and post-task. Three themes identified: (1) becoming aware of the simulation context, (2) defining comfort and confidence in context, and (3) developing models of learning.
Edward Casswell (2015)	British Journal of Ophthalmology	Quantitative / Yes / No	32	100.00%	"this is the first study to investigate whether ophthalmology trainees are able to assess their own competence in cataract surgery when judged with a validated assessment tool (OSACSS)5 and whether this judgement improves as they progress through training, in keeping with the Conscious Competence Learning Model."	71% / -	The Conscious Competence Learning Model	self-assessment	Procedural	Concurrent post-task	Self-observation, Self-judgement	Measured	Characterized: Influence of years of training on self-monitoring. Measured: Self-rating compared with expert rating	Accuracy of self-judgment increased with years in training. Most senior trainees had moderate or greater agreement with expert, compared with a minority of junior trainees.

Claire Daley (2016)	Gastrointestinal Endoscopy	Quantitative / Yes / Yes	33	63.60%	"The aim of this study was to compare an in-classroom training session with a self-directed training session. We sought to compare the 2 groups in their accuracy of predicting HGD/CA in patients with BE, the percentage of high-confidence predictions, and their accuracy in high-confidence predictions."	78% / -	None	Confidence	Interpretation	Concurrent during task	Self-judgement	Measured, Characterized	Characterized: Influence of teaching method on self-monitoring (self-directed v. traditional) Measured: Self-rating of confidence with accuracy of diagnosis.	Both groups had similar alignment of self-rated confidence and correct diagnosis. Participants taught traditionally rated their confidence significantly higher.
Catherine de Blacam (2012)	The American Journal of Surgery	Quantitative / No / No	216	100.00%	"we sought to examine the level of self-awareness in first- and second-year residents."	67% / -	None	Self-assessment, Self-Awareness, Self-Prediction	Procedural	Concurrent pre- and post-task	Self-judgement	Measured	Measured: Pre-task (predicted performance) and post-task self-rating compared with expert post-task rating	Poor correlation between trainee pre-task self-monitoring and expert rating. Moderate correlation of post-task self-monitoring and expert rating. Age (i.e., older), years in training, and nationality significantly influenced self-monitoring accuracy.
Sandeep Ganni (2017)	Journal of Surgical Education	Quantitative / Yes / No	60	100.00%	"This study aimed to determine whether implementing a self-assessment training tool in a validated laparoscopic surgical skills course will improve the accordance between self- and expert assessment."	68% / -	None	Self-assessment	Procedural	Concurrent post-task	Self-judgement	Taught, Measured	Taught: One group of participants trained to self-monitor using instrument Measured: Self-rating compared with expert rating	Participants trained to use the instrument (intervention) had better agreement with experts. Additionally, participants trained to use instrument had better performance.
Luke A Devine (2015)	Simulation in Healthcare	Quantitative / Yes / Yes	39	100.00%	"we compared the educational and cost effectiveness of DSRL with instructor-regulated learning (IRL), using a simulation-based mastery learning model."	89% / -	Mastery Learning Theory, Directed Self-Regulated Learning Theory	Debriefing, Self-regulated	Team Based Care	Concurrent post-task	Self-Judgement, Self-Reaction	Encouraged	Encouraged: One group instructed to engage in self-monitoring, as an element of self-regulated learning, during and after performance.	The group that encourage to engage in self-monitoring performed as well as the group that receive tradition classroom instruction.

Ann W Evans (2007)	Medical Education	Quantitative / No / No	38	Unknown	"We aimed to look at this question further by comparing the results of self-assessment by trainee surgeons with the results obtained by peer assessment of the same procedure."	67% / -	None	Self-Assessment	Procedural	Concurrent post-task	Self-judgement	Meausred	Measured: Self-ratings compared with expert and peer ratings.	Trainee self-ratings had moderate agreement (and significant difference) compared to expert. Self-ratings had high agreement with peer ratings. Trainees with worse performance seemed to have less accurate self-ratings.
James E Allen (2018)	Journal of Clinical Gastroenterology	Quantitative / Yes / Yes	20	60.00%	"The primary aim of our study was to test a group of medical students, residents, and gastroenterology fellows in diagnosing colon polyp histology with NBI using NICE criteria by comparing the efficacy of in-classroom versus self-directed teaching methods for diagnosing colorectal polyps with NBI using the NICE criteria."	68.1% / -	None	Self-assessment	Interpretation	Concurrent during task	Self-judgement	Measured	Characterized: Influence of content instruction (self-directed v. traditional) on self-monitoring. Meausred: Self-rating of confidence compared with histologic diagnosis.	Alignment of confidence and accuracy of diagnosis was similar in the two groups, with a high degree of alignment.
Lisa M Schilling (2005)	Academic Medicine	Quantitative / No / No	43	100.00%	"In our study, we prospectively examined how formulating and answering a patient-specific clinical question affected residents' patient care decisions with internal medicine residents formulating the questions, retrieving the answers, and applying the information to their patients."	39% / -	None	None	Management	Concurrent post-task	Self-Judgement, Self-Reaction	Encouraged	Encouraged: Trainees given blocked-time immediately following a clinic appointment	Intervention had high trainee satisfaction with increased confidence and perceived improvement in communication, medical knowledge, and ability to care for patients with similar problems in the future.
Adam P Sawatsky (2018)	Academic Medicine	Qualitative / No / No	46	100.00%	"the purpose of this study was to explore residents' perceptions of the role that faculty members play in the promotion and support of resident SDL, to better characterize the SDL process in the clinical learning environment."	- / 92.9%	Self-Directed Learning Theory	Self-monitoring	None	None	Self-Observation	Characterized	Characterized: Focus group to understand experiences with self-directed learning, including self-monitoring	Experts acted as triggers for and calibration of self-monitoring

Thomas Geeraerts (2017)	Anaesthesia Critical Care & Pain Medicine	Quantitative / No / No	27	100.00%	"We aimed at describing the intensity of psychological and physical stress induced by simulation training sessions, using both validated surveys as well as biochemical marker of stress, in a group of anaesthesia and critical care residents, and to assess its impact on technical and non-technical performances."	58% / -	None	Self-Assessment	Team Based Care	Concurrent pre- and post-task	Self-Judgement	Measured	Measured: Self-rated stress compared with biologic markers of stress	Self-rated stress correlated with biologic markers of stress (e.g., salivary amylase)
Sandeep Ganni (2017)	Surgical Endoscopy	Quantitative / No / No	35	100.00%	"The aim of this study was to assess the validity of using self-assessment within the Laparoscopic Surgical Skills curriculum (an initiative of the European Association of Endoscopic Surgery)."	77% / -	None	Self-Assessment	Procedural	Concurrent post-task	Self-judgement	Measured	Measured: Self-ratings compared with expert ratings	Strong correlation between self-ratings and expert ratings in all but one category of evaluation instrument
Roxana Geoffrion (2019)	International Urogynecology Journal	Quantitative / Yes / Yes	46	100.00%	Our objectives in the current study were to evaluate whether each module improved self-confidence and satisfaction during performance of the index procedure in the real operating room (OR)."	82% / -	None	self-confidence	Procedural	Concurrent post-task	Self-judgement, Self-Reaction	Characterized	Characterized: Influence of teaching method on self-monitoring (self-directed v. expert coaching)	Self-rated confidence and satisfaction were higher for trainees with expert coaching compared to self-directed learning for one of three procedures (and no difference for two).
Anne Gaunt (2018)	Academic Medicine	Qualitative / No / No	42	Unknown	"Can a self-motives framework of feedback-seeking behavior explain why surgical trainees choose to seek feedback, in the context of WBA, within the clinical workplace? And do contextual factors affect the circumstances in which specific self-motives predominate?"	- / 91%	Self-Motives Model of Feedback	None	Unknown	Concurrent post-task	Self-judgement, Self-Reaction	Characterized	Characterized: Focus groups to understand motives of feedback-seeking	Calibration of self-monitoring, predominately self-judgment, a motivation for seeking feedback. Self-monitoring promotes feedback seeking for performance improvement.

Michael L Green (2009)	Journal of Continuing Education in the Health Professions	Quantitative / No / No	27	14.80%	N/A	52% / -	None	reflection-in-action, reflection-on-action	Management, Interpretation	Concurrent post-task	Self-judgement, Self-Reaction	Encouraged, Characterized	Encouraged: Self-monitoring incorporated as a mechanism for maintenance of certification through logged clinical portfolio Characterized: Influence on medical knowledge and clinical practice	Low rate (36%) of participation. Of the participants, high rates of perceived gains in medical knowledge (84%) and satisfaction (84%). Moderate rate (46%) of perceived change in clinical practice.
Michael L Green (2000)	American Journal of Medicine	Quantitative / No / No	64	100.00%	"our objective was to determine the frequency, characteristics, and pursuit of residents' medical information needs in clinic by interviewing them immediately after each patient encounter."	58% / -	None	None	Management, Interpretation	Concurrent post-task	Self-Judgement, Self-Reaction	Characterized, Encouraged	Characterized: Interview to understand self-monitoring Encouraged: Probing questions to promote self-monitoring	Probing questions seemed to increase self-monitoring, predominately self-reflection with actionable follow-up occuring in 29% of participants.
Ryan Graddy (2018)	Journal of Graduate Medical Education	Quantitative / No / No	46	100.00%	"We sought to evaluate the differences between resident self-assessment and faculty observation and to assess the impact of coaching feedback on goal setting and achievement in an academic general internal medicine (GIM) practice."	52% / -	None	self-assessment	Communication	Concurrent post-task	self-observation, self-Reaction	Taught, Measured	Taught: Coaching, focused on calibration of self-monitoring Measured: Self-ratings compared with expert ratings	Ratings between trainees and faculty similar, but degree of agreement not reported. Trainees had high satisfaction from coaching of self-monitoring (98%) with high rate (90%) of perceived change in practice.
Andrew J Hale (2016)	Academic Medicine	Quantitative / Yes / No	78	100.00%	"we evaluate the effects of self- and peer-review on LR follow-up among house officers in an internal medicine residency training program. We compare residents assigned to peer-review, self-review, or peer- + self-review versus one another and versus historical controls."	78% / -	None	self-review	Communication	Concurrent post-task	Self-Observation	Taught, Characterized, Measured	Characterized: Influence of self-monitoring with or without peer evaluations Taught: Chart audits Measured: Quality of clinical documentation	Self-monitoring alone did not improve documentation. Self-monitor combined with peer evaluation significantly improved quality of clinical documentation.

Andrew B Rosenkrantz (2017)	American Journal of Roentgenology	Quantitative / Yes / Yes	6	100.00%	"In this study, our aim was to evaluate the roles of self-directed learning and continu- al feedback on the learning curve for tumor detection by novice readers of prostate MRI."	76% / -	None	Confidence, Self-directed	Interpretation	Concurrent during task	Self-Judgement, Self-Reaction	Characterized	Characterized: Influence of self-monitoring on interpretation accuracy. Influence of feedback on self-rated confidence.	Accuracy of interpretation improved in self-monitoring with and without feedback. Self-rated confidence increased more in group that received feedback.
Fadi Rzouq (2016)	Endoscopy	Quantitative / Yes / Yes	18	56.00%	"The primary aim of our study was to test two dif- ferent methods for teaching the previously vali- dated criteria by comparing the efficacy of in- class didactic teaching with self-directed teaching in the diagnosis of neoplasia in Barrett's esopha- gus using pCLE among medical trainees. The sec- ondary aim of this study was to determine the learning curve among medical trainees in diag- nosing Barrett's esophagus- associated neoplasia using pCLE."	76% / -	None	Confidence	Interpretation	Concurrent during task	Self-judgement	Characterized, Measured	Characterized: Influence of being able to ask questions to experts on self-monitoring Measured: Self-rating of confidence compared to histopathologic diagnosis	The group that was able to ask questions had greater alignment of self-rated confidence and correct diagnosis.
Annie T Sadosty (2011)	The Journal of Emergency Medicine	Quantitative / No / No	18	100.00%	"1) Determine the accuracy of EM resident perfor- mance self-assessment after a simulation-based en- counter. 2) Compare the ability of low- scoring and high- scoring EM residents to evaluate their performance in a simulation-based assessment exercise. 3) Determine if video-assisted performance review improves accuracy of EM resident self-assessment."	62% / -	None	Self-Assessment	Team Based Care, Management, Communication	Concurrent and retrospective post-task	Self-Judgement, Self-Observation	Measured	Characterized: Influence of concurrent v. retrospective self-judgment Measured: Concurrent and retrospective self-rating compared with expert rating of performance	There was no significant difference in concurrent and retrospective self-ratings of performance. Higher performing residents had higher alignment of self-rating with expert rating. Overall, there was strong agreement between self- and expert ratings.

Michael A Scaffidi (2018)	Gastrointestinal Endoscopy	Quantitative / Yes / No	40	Unknown (52.5-75%)	"this study aimed to evaluate the accuracy of novice, intermediate, and experienced endoscopists in assessing their own competence in performing clinical colonoscopy procedures."	82% / -	None	self-assessment	Procedural	Concurrent post-task	Self-Judgement	Measured	Measured: Self-rating compared with expert rating	Self-ratings of performance had moderate correlation with expert ratings. Participants with greater procedure experience seemed to have higher correlation and be more likely to underestimate their performance (versus overestimation for novices).
Joseph R Schneider (2007)	The American Journal of Surgery	Quantitative / No / No	9	100.00%	"We chose in this study to test the correlation between faculty raters and resident self-rating on the PAME [Patient Assessment and Management Examination] examination."	64% / -	None	self-assessment	Management	Retrospective post-task	Self-judgement	Measured	Measured: Self-rating compared with expert rating	Significant (albeit weak) correlation between self-ratings and expert ratings for one domain of performance. No correlation among the other domains rated.
Mitchell B Alameddine (2015)	Journal of Surgical Education	Quantitative / No / No	Unknown (111 cases)	100.00%	"The purpose of this study is to examine and compare attendings' and residents' perceptions of laparoscopic skills, comfort level, and degree of operative teaching."	62% / -	None	self-assessment	Procedural	Concurrent post-task	Self-Judgement	Measured	Measured: Self-rating compared with expert rating	Experts consistently trainee performance higher than the trainees themselves.

Friedman, Charles P (2005)	Journal of General Internal Medicine	Quantitative / Yes / No	215	33.00%	"This study addresses a question central to the potential utility and success of clinical decision support. If clinicians' openness to external advice hinges on their confidence in their assessments based on personal knowledge, how valid are these perceptions?" Focused on the following questions: In internal medicine, what is the relationship between clinicians' confidence in their diagnoses and the correctness of these diagnoses? 2. Does the relationship between confidence and correctness depend on clinicians' levels of experience ranging from medical student to attending physician? 3. To the extent that confidence and correctness are mismatched, do clinicians tend toward overconfidence or underconfidence, and does this tendency depend on level of clinical experience?	78% / -	None	Confidence, Need for assistance	Interpretation	Concurrent post-task	Self-judgement	Measured	Measured: Self-rated need for assistance compared with accuracy of diagnosis	Significant correlation between diagnostic accuracy and self-rated need for asistance.

Cavalcanti, Rodrigo B. (2014)	Academic Medicine	Quantitative / Yes / Yes (in 1 of 3 experiments)	185	100.00%	"our overall goals were to document the relationship between certainty and accuracy in the setting of physical diagnosis and to explore the effects of data consistency on measures of certainty. Specifically, we aim to address three related questions: What is the variability in responses for diagnostic certainty in cardiac physical diagnosis among postgraduate trainees in medicine across three experimental conditions? How does the variability in diagnostic certainty relate to diagnostic accuracy? Do experimental designs that provide inconsistent clinical data lead to changes in the relationship between reported certainty and accuracy?"	77% / -	None	Confidence, Probability	Interpretation	Concurrent post-task	Self-judgement	Characterized, Measured	Characterized: Influence of distractors on self-monitoring Measured: Self-rated confidence compared with accuracy of diagnosis.	Without any distractors, there was moderate correlation between diagnostic accuracy and self-rated confidence. Distractors (e.g., discordant findings, biased comment) eliminated the correlation of self-rated confidence and diagnostic accuracy.
Jamshidi, Ramin (2009)	Journal of American College of Surgery	Quantitative / Yes / Yes	14	100.00%	"We sought to determine whether such self-assessment would benefit junior surgical trainees in the development of videoscopic skills."	91% / -	None	Review, Self-assessment	Procedural	Retrospective post-task	Self-observation, self-Reaction	Characterized	Characterized: Influence of self-monitoring opportunity on performance	Participants with greater opportunity for retrospective self-monitoring (i.e., video review) had greater improvement on performance of pre-post task compared to participants without video review.
Bonrath, Esther (2015)	Annals of Surgery	Quantitative / Yes / Yes	18	100.00%	"The aim of the present study was to assess the effectiveness of an instructional approach, “comprehensive surgical coaching,” using structured feedback, debriefing, and behavioral modeling as a means to enhance experiential learning in the OR and improve surgical skill and clinical safety."	97% / -	Experiential Learning Theory	Self-assessment, Reflection	Procedural	Concurrent and retrospective post-task	Self-observation, Self-Reaction, Self-Judgement	Characterized, Measured	Characterized: Influence of performance coaching on accuracy of self-judgment of performance Measured: Self-ratings compared with expert ratings of performance.	Trainees that received performance coaching with video review had strong correlation of their self-ratings with expert ratings of performance. Trainees without coaching had no correlation of their self-ratings with expert ratings.

Ward, Mylène (2002)	American Journal of Surgery	Quantitative / No / No	27	100.00%	"What is the accuracy of self-assessment for the performance of a laparoscopic operation? (2) Do interventions—self-observation of videotaped performance, and review of benchmark performances—lead to an improvement in self-assessment ability?"	76% / -	None	self-assessment, self-evaluation	Procedural	Concurrent and retrospective post-task	Self-judgement	Measured	Characterized: Influence of time of self-judgment (concurrent v. retrospective) on accuracy. Measured: Self-ratings compared with expert ratings	Concurrent self-rating had moderate correlation with expert ratings. Retrospective self-rating had higher correlation, which was also similar to the inter-rated reliability among experts.

Appendix 4: Task analysis example (Chapter 3)

- Plan interpretive approach
    - Assess cardinal views
        - Good visualization of structures
        - Gestalt of muscular function (squeeze)
        - Gestalt of valvular function (fully open, close)
        - Gestalt of chamber size
        - Gestalt of wall thickness
    - Review patient data
        - Consult question
        - Age and available past medical history
        - Comparison images/reports
    - Synthesize above to prioritize interpretation allocation based on following:
        - What measurements/views do I need to spend extra time on to answer consult question?
        - What measurements/views do I need to spend extra time on to answer adequately investigate areas of concern based on gestalt?

- Anatomical approach to detailed echocardiography interpretation
    - Overview of approach
        - Recall gestalt. If high certainty and not consult question:
            - Make preferred measurement and
                - If aligns with gestalt, document
                - If does not align with gestalt, perform additional measurements for triangulation and/or view prior, then document
                - If data not aligning appropriately, ask colleague for help and/or view prior
                - If unable to make preferred measurements, document poor data
        - Recall gestalt. If low certainty and not consult question
            - Make more than 1 measurement for triangulation
                - If multiple measurements align, document
                - If multiple measurements misaligned, perform additional measurements and/or view prior, then document with caveat
                - If multiple measurements misaligned and remain uncertain, ask colleague
                - If unable to make multiple measurements, caveat and/or document poor data and/or view prior; recommend alternative modality
        - Recall gestalt. If high certainty and consult question

        - Make more than 1 measurement for triangulation
            - If multiple measurements and gestalt align, document
            - If multiple measurements align, but misalign with gestalt, document measurements
            - If multiple measurements misalign and variable alignment with gestalt, perform additional measurements and/or ask for independent gestalt, document measurements aligning with gestalt and caveat with possible recommendation of alternative diagnostic modality
            - If unable to make multiple measurement, caveat and/or document poor data, recommending alternative diagnostic modality
    - Recall gestalt. If low certainty and consult question
        - Make more than 1 measurement for triangulation
            - If multiple measurements and gestalt align, document
            - If multiple measurements align, but misalign with gestalt, document measurements
            - If multiple measurements misalign and variable alignment with gestalt, perform additional measurements and/or ask for independent measurements, document strongest measurements, independent of gestalt, and caveat with possible recommendation of alternative diagnostic modality
            - If unable to make multiple measurement, caveat and/or document poor data, recommending alternative diagnostic modality
- Chambers size
    - RA
    - LA
    - RV
    - LV
    - Changes from prior
- Chamber function
    - RV
    - LV
    - Changes from prior
- Valvular appearance
    - Thickness
    - Leaflet
    - Reconcile with gestalt
- Valvular function
    - Movement of leaflets
    - Area of valve (assess for stenosis)

            - Size
            - Gradient
        - Regurgitation
    - Effusion
        - Presence
        - Impact on function
        - Changes from prior
    - Great vessels
        - Size
        - Dissection

References

Abimanyi-Ochom, J., Bohingamu Mudiyanselage, S., Catchpool, M., Firipis, M., Wanni Arachchige Dona, S., & Watts, J. J. (2019). Strategies to reduce diagnostic errors: A systematic review. *BMC Medical Informatics and Decision Making*, *19*(1), 7–11. https://doi.org/10.1186/s12911-019-0901-1

ACGME & ABIM. (2013). The Internal Medicine Milestone Project. *The Internal Medicine Milestone Project*, *July*, 1–28. https://doi.org/10.4300/JGME-06-01s1-05

ACGME & ABS. (2015). *The General Surgery Milestone Project*. *July*.

Allen, J. E., Vennalaganti, P., Gupta, N., Hornung, B., Choudhary, A., Titi, M., Alsop, B. R., Lim, D., & Sharma, P. (2018). Randomized Controlled Trial of Self-directed Versus In-Classroom Education of Narrow Band Imaging in Diagnosing Colorectal Polyps Using the NICE Criteria. *Journal of Clinical Gastroenterology*, *52*(5), 413–417. https://doi.org/10.1097/MCG.0000000000000791

Archer, J. C. (2010). State of the science in health professional education: Effective feedback. *Medical Education*, *44*(1), 101–108. https://doi.org/10.1111/j.1365-2923.2009.03546.x

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Artino, A. R., Brydges, R., & Gruppen, L. D. (2015). Self-regulated learning in healthcare profession education: Theoretical perspectives and research methods. *Researching Medical Education*, 155–166. https://doi.org/10.1002/9781118838983.ch14

Artino, A. R., Cleary, T. J., Dong, T., Hemmer, P. A., & Durning, S. J. (2014). Exploring clinical reasoning in novices: A self-regulated learning microanalytic assessment approach. *Medical Education*, *48*(3), 280–291. https://doi.org/10.1111/medu.12303

Atman, C. J., & Bursic, K. M. (1998). Verbal Protocol Analysis as a Method to Document Engineering Student Design Processes. *Journal of Engineering Education*.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215.

Bates, D. W., & Singh, H. (2018). Two Decades Since *To Err Is Human*: An Assessment Of Progress And Emerging Priorities In Patient Safety. *Health Affairs*, *37*(11), 1736–1743. https://doi.org/10.1377/hlthaff.2018.0738

Beatty, P. C., & Willis, G. B. (2007). Research Synthesis: The Practice of Cognitive Interviewing. *Public Opinion Quarterly*, *71*(2), 287–311. https://doi.org/10.1093/poq/nfm006

Bennett-Levy, J. (2006). Therapist Skills: A Cognitive Model of their Acquisition and Refinement. *Behavioural and Cognitive Psychotherapy*, *34*(1), 57–78. https://doi.org/10.1017/S1352465805002420

Boet, S., Bould, M. D., Bruppacher, H. R., Desjardins, F., Chandra, D. B., & Naik, V. N. (2011). Looking in the mirror: Self-debriefing versus instructor debriefing for simulated crises*: *Critical Care Medicine*, *39*(6), 1377–1381. https://doi.org/10.1097/CCM.0b013e31820eb8be

Bond, W. F., Deitrick, L. M., Arnold, D. C., Kostenbader, M., Barr, G. C., Kimmel, S. R., & Worrilow, C. C. (2004). Using Simulation to Instruct Emergency Medicine Residents in

Cognitive Forcing Strategies: *Academic Medicine*, *79*(5), 438–446. https://doi.org/10.1097/00001888-200405000-00014

Bonrath, E. M., Dedy, N. J., Gordon, L. E., & Grantcharov, T. P. (2015). Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial. *Annals of Surgery*, *262*(2), 205–212. https://doi.org/10.1097/SLA.0000000000001214

Bounds, R., Bush, C., Aghera, A., Rodriguez, N., Stansfield, R. B., & Santen, S. A. (2013). Emergency medicine residents' self-assessments play a critical role when receiving feedback. *Academic Emergency Medicine*, *20*(10), 1055–1061. https://doi.org/10.1111/acem.12231

Boyce, C., & Neale, P. (2006). *CONDUCTING IN-DEPTH INTERVIEWS:* 17.

Brydges, R., Dubrowski, A., & Regehr, G. (2010). A New Concept of Unsupervised Learning: Directed Self-Guided Learning in the Health Professions: *Academic Medicine*, *85*, S49–S55. https://doi.org/10.1097/ACM.0b013e3181ed4c96

Brydges, R., Hatala, R., & Mylopoulos, M. (2016). Examining Residents' Strategic Mindfulness During Self-Regulated Learning of a Simulated Procedural Skill. *Journal of Graduate Medical Education*, *8*(3), 364–371. https://doi.org/10.4300/JGME-D-15-00491.1

Burbach, B., Barnason, S., & Thompson, S. A. (2015). Using "Think Aloud" to Capture Clinical Reasoning during Patient Simulation. *International Journal of Nursing Education Scholarship*, *12*(1), 1–7. https://doi.org/10.1515/ijnes-2014-0044

Carcary, M. (2009). *The Research Audit Trial—Enhancing Trustworthiness in Qualitative Inquiry*. *7*(1), 15.

Casswell, E. J., Salam, T., Sullivan, P. M., & Ezra, D. G. (2016). Ophthalmology trainees' self-assessment of cataract surgery. *British Journal of Ophthalmology*, *100*(6), 766–771. https://doi.org/10.1136/bjophthalmol-2015-307307

Charters, E. (2003). The Use of Think-aloud Methods in Qualitative Research An Introduction to Think-aloud Methods. *Brock Education*, *12*(2), 68–82. https://doi.org/10.1093/ije/29.4.773

Cho, K. K., Marjadi, B., Langendyk, V., & Hu, W. (2017). The self-regulated learning of medical students in the clinical environment—A scoping review. *BMC Medical Education*, *17*(1), 1–13. https://doi.org/10.1186/s12909-017-0956-6

Cleary, T. J., Durning, S. J., Gruppen, L. D., Hemmer, P. A., & Artino, A. R. (2013). Self-Regulated Learning in Medical Education. In *Oxford Textbook of Medical Education*. Oxford University Press.

Cleary, T. J., & Sandars, J. (2011). Assessing self-regulatory processes during clinical skill performance: A pilot study. *Medical Teacher*, *33*(7). https://doi.org/10.3109/0142159X.2011.577464

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Mastery learning for health professionals using technology-enhanced simulation: A systematic review and meta-analysis. *Academic Medicine*, *88*(8), 1178–1186. https://doi.org/10.1097/ACM.0b013e31829a365d

Cook, D. A., & Reed, D. A. (2015). Appraising the Quality of Medical Education Research Methods: The Medical Education Research Study Quality Instrument and the Newcastle-

Ottawa Scale-Education. *Academic Medicine*, *90*(8), 1067–1076. https://doi.org/10.1097/ACM.0000000000000786

Corbin, J., & Morse, J. M. (2003). The Unstructured Interactive Interview: Issues of Reciprocity and Risks when Dealing with Sensitive Topics. *Qualitative Inquiry*, *9*(3), 335–354. https://doi.org/10.1177/1077800403009003001

Cowan, N. (2014). Working Memory Underpins Cognitive Development, Learning, and Education. *Educational Psychology Review*, *26*(2), 197–223. https://doi.org/10.1007/s10648-013-9246-y

Croft, H., Gilligan, C., Rasiah, R., Levett-Jones, T., & Schneider, J. (2018). *Thinking in Pharmacy Practice: A Study of Community Pharmacists' Clinical Reasoning in Medication Supply Using the Think-Aloud Method*. 14.

Crommelinck, M., & Anseel, F. (2013). Understanding and encouraging feedback-seeking behaviour: A literature review: Feedback-seeking behaviour: a review. *Medical Education*, *47*(3), 232–241. https://doi.org/10.1111/medu.12075

Davis, D. A., Mazmanian, P. E., Fordis, M., Harrison, R. V., Thorpe, K. E., Perrier, L., D.A., D., P.E., M., M., F., Van, H. R., K.E., T., & L., P. (2006). Accuracy of Physician Self-assessment Compared with Observed Measures of Competence. *JAMA - Journal of the American Medical Association*, *296*(9), 1094–1102.

Day, S., Norcini, J., Webster, G., Viner, E., & Chirico, A. (1988). The effect of changes in medical knowledge on examination performance at the time of recertification. *Proceeding of the Annual Conference of Research in Medical Education*, *27*, 139–144.

De Blacam, C., O'Keeffe, D. A., Nugent, E., Doherty, E., & Traynor, O. (2012). Are residents accurate in their assessments of their own surgical skills? *American Journal of Surgery*, *204*(5), 724–731. https://doi.org/10.1016/j.amjsurg.2012.03.003

Durning, S. J., & Artino, A. R. (2011). Situativity theory: A perspective on how participants and the environment can interact: AMEE Guide no. 52. *Medical Teacher*, *33*(3), 188–199. https://doi.org/10.3109/0142159X.2011.550965

Durning, S. J., Artino, A. R., Beckman, T. J., Graner, J., Van Der Vleuten, C., Holmboe, E., & Schuwirth, L. (2013). Does the think-aloud protocol reflect thinking? Exploring functional neuroimaging differences with thinking (answering multiple choice questions) versus thinking aloud. *Medical Teacher*, *35*(9), 720–726. https://doi.org/10.3109/0142159X.2013.801938

Durning, S. J., Graner, J., Artino, A. R., Pangaro, L. N., Beckman, T., Holmboe, E., Oakes, T., Roy, M., Riedy, G., Capaldi, V., Walter, R., van der Vleuten, C., & Schuwirth, L. (2012). Using functional neuroimaging combined with a think-aloud protocol to explore clinical reasoning expertise in internal medicine. *Military Medicine*, *177*(SUPPL.1), 72–78. https://doi.org/10.7205/milmed-d-12-00242

Duschl, K. C., Gramß, D., Obermeier, M., & Vogel-Heuser, B. (2015). Towards a taxonomy of errors in PLC programming. *Cognition, Technology and Work*, *17*(3), 417–430. https://doi.org/10.1007/s10111-014-0307-x

Eccles, D. W., & Arsal, G. (2017). The think aloud method: What is it and how do I use it? *Qualitative Research in Sport, Exercise and Health*, *9*(4), 514–531. https://doi.org/10.1080/2159676X.2017.1331501

Edgar, L., McLean, S., Hogan, S. O., Hamstra, S., & Holmboe, E. S. (2020). *The Milestones Guidebook*.

Edwards, R., & Holland, J. (2013). *What is Qualitative Interviewing?* Bloomsbury Academic. https://doi.org/10.5040/9781472545244

Ericsson, K. A. (2002). *Protocol analysis and Verbal Reports on Thinking An updated and extracted version from Ericsson (2002)*. 1–4.

Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, A. M. (Eds.). (2018). *The Cambridge Handbook of Expertise and Expert Performance* (Second). Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215

Ericsson, K. A., & Simon, H. A. (1981). Protocol Analysis. *A Companion to Cognitive Science*, 425–432. https://doi.org/10.1002/9781405164535.ch33

Ericsson, K. A., & Simon, H. A. (1984). *Protocol Analysis: Verbal reports as data*. The MIT Press.

Eva, K. W., & Regehr, G. (2005). Self-Assessment in the Health Professions: A Reformulation and Research Agenda. *Academic Medicine*, *80*(Supplement), S46–S54. https://doi.org/10.1097/00001888-200510001-00015

Eva, K. W., & Regehr, G. (2009). "I'll Never Play Professional Football" and Other Fallacies of Self-Assessment. *Journal of Continuing Education in the Health Professions*, *28*(1), 14–19. https://doi.org/10.1002/chp

Eva, K. W., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education*, *16*(3), 311–329. https://doi.org/10.1007/s10459-010-9263-2

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research*, *3*(4), 430–441.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do Procedures for Verbal Reporting of Thinking Have to Be Reactive? A Meta-Analysis and Recommendations for Best Reporting Methods. *Psychological Bulletin*, *137*(2), 316–344. https://doi.org/10.1037/a0021663

Francis, N. K., Hanna, G. B., Cresswell, A. B., Carter, F. J., & Cuschieri, A. (2001). The performance of master surgeons on standard aptitude testing. *The American Journal of Surgery*, *182*(1), 30–33. https://doi.org/10.1016/S0002-9610(01)00652-3

Frank, J. R., Snell, L., Sherbino, J., & Royal College of Physicians and Surgeons of Canada. (2015). *CanMEDS 2015 physician competency framework*.

Fuller, S. (2001). *Positivism, History of* (N. J. Smelser & P. B. B. T.-I. E. of the S. & B. S. Baltes, Eds.; pp. 11821–11827). Pergamon. https://doi.org/10.1016/B0-08-043076-7/00084-X

Gagné, R. M., & Smith, E. C. (1962). A study of the effects of verbalization on problem solving. *Journal of Experimental Psychology*, *63*(1), 12–18. https://doi.org/10.1037/h0048703

Ganni, S., Botden, S. M. B. I., Schaap, D. P., Verhoeven, B. H., Goossens, R. H. M., & Jakimowicz, J. J. (2018). "Reflection- Before -Practice" Improves Self-Assessment and End-Performance in Laparoscopic Surgical Skills Training. *Journal of Surgical Education*, *75*(2), 527–533. https://doi.org/10.1016/j.jsurg.2017.07.030

Ganni, S., Chmarra, M. K., Goossens, R. H. M., & Jakimowicz, J. J. (2017). Self-assessment in laparoscopic surgical skills training: Is it reliable? *Surgical Endoscopy*, *31*(6), 2451–2456. https://doi.org/10.1007/s00464-016-5246-6

Gardin, F. A. (n.d.). *The "Think-Aloud" Method to Promote Student Modeling of Expert Thinking*. 2.

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-Based Assessments as Social Judgments: Rethinking the Etiology of Rater Errors. *Academic Medicine*, *86*(10), S1–S7. https://doi.org/10.1097/ACM.0b013e31822a6cf8

Göransson, K. E., Ehrenberg, A., Ehnfors, M., & Fonteyn, M. (2007). An effort to use qualitative data analysis software for analysing think aloud data. *International Journal of Medical Informatics*, *76*(SUPPL. 2), 270–273. https://doi.org/10.1016/j.ijmedinf.2007.05.004

Graddy, R., Reynolds, S. S., & Wright, S. M. (2018). Coaching Residents in the Ambulatory Setting: Faculty Direct Observation and Resident Reflection. *Journal of Graduate Medical Education*, *10*(4), 449–454. https://doi.org/10.4300/JGME-17-00788.1

Grubbs, M. E., Strimel, G. J., & Kim, E. (2018). Examining design cognition coding schemes for P-12 engineering/technology education. *International Journal of Technology and Design Education*, *28*(4), 899–920. https://doi.org/10.1007/s10798-017-9427-y

Henry, S., LeBreck, D., & Holzemer, W. (1989). The effect of verbalization of cognitive processes on clinical decision making. *Res Nurs Health*, *12*(3), 187–193. https://doi.org/10.1002/nur.4770120309

Hiemstra, R., & Brockett, R. G. (2012). *Reframing the Meaning of Self-Directed Learning: An Updated Model*.

Hildebrand, C., Trowbridge, E., Roach, M. A., Sullivan, A. G., Broman, A. T., & Vogelman, B. (2009). Resident self-assessment and self-reflection: University of Wisconsin-Madison's

five-year study. *Journal of General Internal Medicine*, *24*(3), 361–365. https://doi.org/10.1007/s11606-009-0904-1

Holmboe, E. S., Prince, L., & Green, M. (2005). *Teaching and Improving Quality of Care in a Primary Care Internal Medicine Residency Clinic*. *80*(6), 571–577.

Houten-schat, M. A. V., Berkhout, J. J., van Dijk, N., Endedijk, M. D., Jaarsma, A. D. C., Diemers, A. D., van Houten-Schat, M. A., Berkhout, J. J., van Dijk, N., Endedijk, M. D., Jaarsma, A. D. C., & Diemers, A. D. (2018). Self-regulated learning in the clinical context: A systematic review. *Medical Education*, *52*(10), 1008–1015. https://doi.org/10.1111/medu.13615

Ilgen, J. S., Bowen, J., de Bruin, A., Regehr, G., & Teunissen, P. (2020). "I Was Worried About the Patient, But I Wasn't Feeling Worried." *Academic Medicine*, *Publish Ah*. https://doi.org/10.1097/acm.0000000000003634

Jamshidi, R., LaMasters, T., Eisenberg, D., Duh, Q.-Y., & Curet, M. (2009). Video Self-Assessment Augments Development of Videoscopic Suturing Skill. *Journal of the American College of Surgeons*, *209*(5), 622–625. https://doi.org/10.1016/j.jamcollsurg.2009.07.024

Kelley, T. R., Capobianco, B. M., & Kaluf, K. J. (2015). Concurrent think-aloud protocols to assess elementary design students. *International Journal of Technology and Design Education*, *25*(4), 521–540. https://doi.org/10.1007/s10798-014-9291-y

Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE Guide No. 131. *Medical Teacher*, *42*(8), 846–854. https://doi.org/10.1080/0142159X.2020.1755030

Kilsdonk, E. (2016). Uncovering healthcare practitioners' information processing using the think-aloud method: From paper-based guideline to clinical decision support system. *International Journal of Medical Informatics*, 10.

Kim, Y.-M. (2002). Impact of supervision and self-assessment on doctor-patient communication in rural Mexico. *International Journal for Quality in Health Care*, *14*(5), 359–367. https://doi.org/10.1093/intqhc/14.5.359

Kleynen, M., Moser, A., Haarsma, F. A., Beurskens, A. J., & Braun, S. M. (2017). Physiotherapists use a great variety of motor learning options in neurological rehabilitation, from which they choose through an iterative process: A retrospective think-aloud study. *Disability and Rehabilitation*, *39*(17), 1729–1737. https://doi.org/10.1080/09638288.2016.1207111

Kogan, J. R., & Holmboe, E. S. (2017). Direct Observation. In E. S. Holmboe, S. J. Durning, & R. E. Hawkins (Eds.), *Practical Guide to Evaluation of Clinical Competence* (pp. 66–90). Elsevier.

Kolb, D. (1984). Experiential Learning: Experience as the source of learning and development. In *Journal of Business Ethics* (Vol. 1).

Könings, K. D., van Berlo, J., Koopmans, R., Hoogland, H., Spanjers, I. A. E., ten Haaf, J. A., van der Vleuten, C. P. M., & van Merriënboer, J. J. G. (2016). Using a Smartphone App and Coaching Group Sessions to Promote Residents' Reflection in the Workplace: *Academic Medicine*, *91*(3), 365–370. https://doi.org/10.1097/ACM.0000000000000989

Koriat, A. (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.

Koro-Ljungberg, M., Douglas, E. P., Therriault, D., Malcolm, Z., & McNeill, N. (2013). Reconceptualizing and decentering think-aloud methodology in qualitative research. *Qualitative Research*, *13*(6), 735–753. https://doi.org/10.1177/1468794112455040

Leppink, J., van Gog, T., Paas, F., & Sweller, J. (2015). Cognitive load theory: Researching and planning teaching to maximise learning. In J. Cleland & S. J. Durning (Eds.), *Researching Medical Education* (1st ed., pp. 207–218). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118838983.ch18

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, *5*(1), 69. https://doi.org/10.1186/1748-5908-5-69

Lundgrén-Laine, H., & Salanterä, S. (2010). Think-aloud technique and protocol analysis in clinical decision-making research. *Qualitative Health Research*, *20*(4), 565–575. https://doi.org/10.1177/1049732309354278

Mandel, L. S., Goff, B. A., & Lentz, G. M. (2005). Self-assessment of resident surgical skills: Is it feasible? *American Journal of Obstetrics and Gynecology*, *193*(5), 1817–1822. https://doi.org/10.1016/j.ajog.2005.07.080

Maslow, A., Frager, R., & Fadiman, James. (1987). *Motivation and personality* (3rd ed.). Harper and Row.

McConnell, M. M., Regehr, G., Wood, T. J., & Eva, K. W. (2012). Self-monitoring and its relationship to medical knowledge. *Advances in Health Sciences Education*, *17*(3), 311–323. https://doi.org/10.1007/s10459-011-9305-4

McGaghie, W. C. (2008). Research Opportunities in Simulation-based Medical Education Using Deliberate Practice. *Academic Emergency Medicine*, *15*(11), 995–1001. https://doi.org/10.1111/j.1553-2712.2008.00246.x

McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., & Scalese, R. J. (2010). A critical review of simulation-based medical education research: 2003–2009: Simulation-based medical education research 2003–2009. *Medical Education*, *44*(1), 50–63. https://doi.org/10.1111/j.1365-2923.2009.03547.x

Moulton, C. A., Regehr, G., Lingard, L., Merritt, C., & MacRae, H. (2010). "Slowing Down When You Should": Initiators and influences of the transition from the routine to the effortful. *Journal of Gastrointestinal Surgery*, *14*(6), 1019–1026. https://doi.org/10.1007/s11605-010-1178-y

Nelson, T. O. (1990). *Metamemory: A Theoretical Framework and New Findings* (G. H. Bower, Ed.; Vol. 26, pp. 125–173). Academic Press. https://doi.org/10.1016/S0079-7421(08)60053-5

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

O'Brien, B. C., Harris, I. B., Beckman, T. J., Reed, D. A., & Cook, D. A. (2014). Standards for reporting qualitative research: A synthesis of recommendations. *Academic Medicine*, *89*(9), 1245–1251. https://doi.org/10.1097/ACM.0000000000000388

Osborne, A. J., Hawkins, S. C., Pournaras, D. J., Chandratilake, M., & Welbourn, R. (2014). An evaluation of operative self-assessment by UK postgraduate trainees. *Medical Teacher*, *36*(1), 32–37. https://doi.org/10.3109/0142159X.2013.836268

Özdemir, I. E. Y. (2011). Self-regulated learning from a sociocultural perspective. *Education and Science*, *36*(160), 298–308.

Patton, M. Q. (2015). *Qualitative Research & Evaluation Methods* (4th ed.). SAGE Publications, Inc.

Picho, K., Maggio, L. A., & Artino, A. R. (2016). Science: The slow march of accumulating evidence. *Perspectives on Medical Education*, *5*(6), 350–353. https://doi.org/10.1007/s40037-016-0305-1

Pinnock, R., Young, L., Spence, F., Henning, M., & Hazell, W. (n.d.). *Can Think Aloud Be Used to Teach and Assess Clinical Reasoning in Graduate Medical Education?* 4.

Plant, J. L., Corden, M., Mourad, M., O'Brien, B. C., & van Schaik, S. M. (2013). Understanding self-assessment as an informed process: Residents' use of external information for self-assessment of performance in simulated resuscitations. *Advances in Health Sciences Education*, *18*(2), 181–192. https://doi.org/10.1007/s10459-012-9363-2

Pomare, C., Churruca, K., Ellis, L. A., Long, J. C., & Braithwaite, J. (2019). A revised model of uncertainty in complex healthcare settings: A scoping review. *Journal of Evaluation in Clinical Practice*, *25*(2), 176–182. https://doi.org/10.1111/jep.13079

Quick, J. A., Kudav, V., Doty, J., Crane, M., Bukoski, A. D., Bennett, B. J., & Barnes, S. L. (2017). Surgical resident technical skill self-evaluation: Increased precision with training progression. *Journal of Surgical Research*, *218*, 144–149. https://doi.org/10.1016/j.jss.2017.05.070

Ravitz, P., Lancee, W. J., Lawson, A., Maunder, R., Hunter, J. J., Leszcz, M., McNaughton, N., & Pain, C. (2013). Improving Physician–Patient Communication Through Coaching of Simulated Encounters. *Academic Psychiatry*, *37*(2), 87. https://doi.org/10.1176/appi.ap.11070138

Richardson, S., Mishuris, R., O'Connell, A., Feldstein, D., Hess, R., Smith, P., McCullagh, L., McGinn, T., & Mann, D. (2018). *"Think Aloud" and "Near Live" Usability Testing of Two Complex Clinical Decision Support Tools*. 20.

Rosenkrantz, A. B., Ayoola, A., Hoffman, D., Khasgiwala, A., Prabhu, V., Smereka, P., Somberg, M., & Taneja, S. S. (2017). The Learning Curve in Prostate MRI Interpretation: Self-Directed Learning Versus Continual Reader Feedback. *American Journal of Roentgenology*, *208*(3), W92–W100. https://doi.org/10.2214/AJR.16.16876

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition*, *17*(6), 759–769. https://doi.org/10.3758/BF03202637

Rzouq, F., Vennalaganti, P., Pakseresht, K., Kanakadandi, V., Parasa, S., Mathur, S., Alsop, B., Hornung, B., Gupta, N., & Sharma, P. (2015). In-class didactic versus self-directed teaching of the probe-based confocal laser endomicroscopy (pCLE) criteria for Barrett's esophagus. *Endoscopy*, *48*(02), 123–127. https://doi.org/10.1055/s-0034-1393118

Sargeant, J., Armson, H., Chesluk, B., Dornan, T., Eva, K., Holmboe, E., Lockyer, J., Loney, E., Mann, K., & Van Der Vleuten, C. (2010). The processes and dimensions of informed self-assessment: A conceptual model. *Academic Medicine*, *85*(7), 1212–1220. https://doi.org/10.1097/ACM.0b013e3181d85a4e

Scaffidi, M. A., Grover, S. C., Carnahan, H., Khan, R., Amadio, J. M., Yu, J. J., Dargavel, C., Khanna, N., Ling, S. C., Yong, E., Nguyen, G. C., & Walsh, C. M. (2018). Impact of experience on self-assessment accuracy of clinical colonoscopy competence. *Gastrointestinal Endoscopy*, *87*(3), 688-694.e2. https://doi.org/10.1016/j.gie.2017.10.040

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, *26*(2), 145–159. https://doi.org/10.1037/h0030806

Singh, H., & Graber, M. L. (2015). Improving Diagnosis in Health Care—The Next Imperative for Patient Safety. *New England Journal of Medicine*, *373*(26), 2493–2495. https://doi.org/10.1056/NEJMp1508044

Skaugset, L. M., Farrell, S., Carney, M., Wolff, M., Santen, S. A., Perry, M., & Cico, S. J. (2016). Can You Multitask? Evidence and Limitations of Task Switching and Multitasking in Emergency Medicine. *Annals of Emergency Medicine*, *68*(2), 189–195. https://doi.org/10.1016/j.annemergmed.2015.10.003

Standards for assessment and accreditation of primary medical programs by the Australian Medical Council 2012. (2012). Australian Medical Council Limited.

Surry, L. T., Torre, D., & Durning, S. J. (2017). Exploring examinee behaviours as validity evidence for multiple-choice question examinations. *Medical Education*, *51*(10), 1075–1085. https://doi.org/10.1111/medu.13367

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Thomas, A., Lubarsky, S., Durning, S. J., & Young, M. E. (2017). Knowledge Syntheses in Medical Education: Demystifying Scoping Reviews. *Academic Medicine*, *92*(2), 161–166. https://doi.org/10.1097/ACM.0000000000001452

Titchener, E. B. (1912). The Schema of Introspection. *The American Journal of Psychology*, *23*(4), 485–508.

Trajkovski, T., Veillette, C., Backstein, D., Wadey, V. M. R., & Kraemer, B. (2012). Resident self-assessment of operative experience in primary total knee and total hip arthroplasty: Is it accurate? *Canadian Journal of Surgery*, *55*(4 Suppl 2), S153–S157. https://doi.org/10.1503/cjs.035510

Ularntinon, S., & Friedberg, R. D. (2016). The SELF: A Supervisory Tool for Enhancing Residents' Self-Reflective Learning in CBT with Youth. *Academic Psychiatry*, *40*(1), 172–176. https://doi.org/10.1007/s40596-014-0264-y

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The Think Aloud Method: A practical guide to modelling cognitive processes*. Academic Press.

Varpio, L., Day, K., Elliot-Miller, P., King, J. W., Kuziemsky, C., Parush, A., Roffey, T., & Rashotte, J. (2015). The impact of adopting EHRs: How losing connectivity affects clinical reasoning. *Medical Education*, *49*(5), 476–486. https://doi.org/10.1111/medu.12665

Veaudor, M., Gérinière, L., Souquet, P.-J., Druette, L., Martin, X., Vergnon, J.-M., & Couraud, S. (2018). High-fidelity simulation self-training enables novice bronchoscopists to acquire basic bronchoscopy skills comparable to their moderately and highly experienced counterparts. *BMC Medical Education*, *18*(1), 191. https://doi.org/10.1186/s12909-018-1304-1

Vyasa, P., Willis, R. E., Dunkin, B. J., & Gardner, A. K. (2017). Are General Surgery Residents Accurate Assessors of Their Own Flexible Endoscopy Skills? *Journal of Surgical Education*, *74*(1), 23–29. https://doi.org/10.1016/j.jsurg.2016.06.018

Ward, M., MacRae, H., Schlachta, C., Mamazza, J., Poulin, E., Reznick, R., & Regehr, G. (2003). Resident self-assessment of operative performance. *The American Journal of Surgery*, *185*(6), 521–524. https://doi.org/10.1016/S0002-9610(03)00069-2

Watson, J. B. (2009). Is thinking merely the action of language mechanisms? *British Journal of Psychology*, *100*(S1), 169–180. https://doi.org/10.1348/000712608x336095

Wolpaw, T. M., Wolpaw, D. R., & Papp, K. K. (2003). SNAPPS: A Learner-centered Model for Outpatient Education. *Academic Medicine*, *78*(9), 893–898. https://doi.org/10.1097/00001888-200309000-00010

Wouda, J. C., & van de Wiel, H. B. M. (2014). The effects of self-assessment and supervisor feedback on residents' patient-education competency using videoed outpatient consultations. *Patient Education and Counseling*, *97*(1), 59–66. https://doi.org/10.1016/j.pec.2014.05.023

Wu, B. J., Dietz, P. A., Bordley, J., & Borgstrom, D. C. (2009). A Novel, Web-Based Application for Assessing and Enhancing Practice-Based Learning in Surgery Residency. *Journal of Surgical Education*, *66*(1), 3–7. https://doi.org/10.1016/j.jsurg.2008.07.015

Yang, S. C. (2003). Reconceptualizing think-aloud methodology: Refining the encoding and categorizing techniques via contextualized perspectives. *Computers in Human Behavior*, *19*(1), 95–115. https://doi.org/10.1016/S0747-5632(02)00011-0

Young, M. (1993). Instructional Design for Situated Learning. *Educational Technology Research and Development1*, *41*(1), 43–58.

Zimmerman, B. J., & Schunk, D. H. (Eds.). (1989). *Self-Regulated Learning and Academic Achievement*. Springer New York. https://doi.org/10.1007/978-1-4612-3618-4