Original Article

SAGE Open Medicine

SAGE Open Medicine
Volume 4: 1–7

DOI: 10.1177/2050312116682254
smo.sagepub.com
SAGE

Performance of a cognitive load inventory during simulated handoffs: Evidence for validity

John Q Young[1], Christy K Boscardin[2], Savannah M van Dijk[3], Ruqayyah Abdullah[1], David M Irby[2], Justin L Sewell[2], Olle Ten Cate[3] and Patricia S O'Sullivan[2]

Abstract

Background: Advancing patient safety during handoffs remains a public health priority. The application of cognitive load theory offers promise, but is currently limited by the inability to measure cognitive load types.

Objective: To develop and collect validity evidence for a revised self-report inventory that measures cognitive load types during a handoff.

Methods: Based on prior published work, input from experts in cognitive load theory and handoffs, and a think-aloud exercise with residents, a revised Cognitive Load Inventory for Handoffs was developed. The Cognitive Load Inventory for Handoffs has items for intrinsic, extraneous, and germane load. Students who were second- and sixth-year students recruited from a Dutch medical school participated in four simulated handoffs (two simple and two complex cases). At the end of each handoff, study participants completed the Cognitive Load Inventory for Handoffs, Paas' Cognitive Load Scale, and one global rating item for intrinsic load, extraneous load, and germane load, respectively. Factor and correlational analyses were performed to collect evidence for validity.

Results: Confirmatory factor analysis yielded a single factor that combined intrinsic and germane loads. The extraneous load items performed poorly and were removed from the model. The score from the combined intrinsic and germane load items associated, as predicted by cognitive load theory, with a commonly used measure of overall cognitive load (Pearson's $r = 0.83$, $p < 0.001$), case complexity (beta = 0.74, $p < 0.001$), level of experience (beta = −0.96, $p < 0.001$), and handoff accuracy ($r = -0.34$, $p < 0.001$).

Conclusion: These results offer encouragement that intrinsic load during handoffs may be measured via a self-report measure. Additional work is required to develop an adequate measure of extraneous load.

Keywords

Cognitive load, handoffs, measurement, validity

Date received: 5 July 2016; accepted: 24 October 2016

Background

Patient handoffs are associated with medical errors and harm to patients.[1,2] Considerable attention in the literature has been focused on interventions to improve patient safety during handoffs,[3] many of which have been adapted from industries in which transition errors have high consequences.[4] These best practices facilitate information transfer via communication protocols that include structured face-to-face and written sign-out, teamwork, interactive questioning, and distraction-free settings.[3,5] Recent implementation of a handoff bundle in multiple pediatric hospitals yielded improvements in educational and clinical outcomes.[6]

Despite these advances, handoffs remain a significant patient safety challenge. Conceptual work has highlighted

[1]Hofstra Northwell School of Medicine, Zucker Hillside Hospital, Glen Oaks, NY, USA
[2]UCSF School of Medicine, University of California–San Francisco, San Francisco, CA, USA
[3]Utrecht Medical Center, Utrecht, The Netherlands

Corresponding author:
John Q Young, Hofstra Northwell School of Medicine, Zucker Hillside Hospital, 75-59 263rd Street, Kaufman 217A, Glen Oaks, NY 11004, USA.
Email: Jyoung9@northwell.edu

cognitive load theory (CLT) as a framework that may help researchers better appreciate the cognitive mechanisms of handoff errors.[7] Originally developed by Sweller and colleagues[8,9] in the context of studying how students problem-solve, CLT focuses on the implications of limited working memory (WM) for learning. Unlike sensory and long-term memory, WM is not infinite—WM can only actively process (i.e. organize, compare, and contrast) no more than two to four elements at any given moment as suggested by the most recent work in the area.[10,11] Theoretically, when the cognitive load of a handoff exceeds the WM capacity of the learner, errors occur, often in the form of information loss (e.g. drug allergy, critical co-morbidity, relevant history, or current treatments) or distortion (e.g. wrong medication dose, wrong surgical site, or incorrect diagnosis).

CLT understands learning as the construction and automation of schemata.[12] Researchers have differentiated overall cognitive load into three types: intrinsic load (IL) (information processing essential to learning the skill), extraneous load (EL) (information processing induced by suboptimal design of the task or the physical environment), and germane load (GL) (information processing imposed by the learner's deliberate use of cognitive strategies to refine existing schemata and enhance storage in long-term memory).[13] Recent work by Sweller and others has suggested that GL may best be understood as a component of IL rather than a separate type of load.[14,15] In this view, a two-factor model (IL and EL) is preferred. Regardless of whether IL and GL are considered separate constructs, CLT's focus on WM as the bottleneck for learning leads to three instructional strategies: minimize EL, match IL to the developmental stage of the learner, and optimize GL.[12]

Researchers have developed a number of techniques to estimate cognitive load,[16,17] including learner self-rating of effort,[18] response time to a secondary task (e.g. participant's response to a vibration sensation) presented during the primary task,[19,20] observations,[21] and psychophysiological measures (e.g. heart rate variability, pupillary response, and electrical skin conductance).[22] While secondary task and physiological measures allow for the objective measurement of cognitive load dynamically throughout the task in contrast to self-ratings which are more subjective and occur only after the task, researchers most commonly use learner self-rating because it is inexpensive and easy to administer.[23] Paas'[23] single-item self-report measure has been used extensively.[24] While developed as a measure of overall cognitive load, some argue that Paas' Scale may actually measure IL rather than overall load.[20,25] The other most commonly used self-rating instrument is the National Aeronautics and Space Administration Task Load Index (NASA-TLX), a multi-item scale that measures overall mental workload.[26,27]

However, the application of CLT has been limited by the absence of measures that differentiate cognitive load types. Such a measure would help identify the cognitive mechanisms of handoff errors and develop new educational strategies and protocols that modulate IL, EL, and GL in the desired directions. Outside of handoffs, the most promising efforts to measure load types have focused on classroom-based learning (e.g. college statistics)[14,28] and colonoscopy performed by gastroenterology fellows.[29] Both groups have developed instruments with evidence for validity that are promising but not directly applicable to handoffs. Only one published study has reported efforts to develop a handoff-specific measure. This study had mixed results.[30] The IL items did form a single factor, but the EL items performed poorly. In addition, this inventory had only a single item for GL and the study did not use an adequate measure of performance.

As of now, the field has yet to develop a measure of cognitive load types during handoffs that has sufficient evidence for validity to warrant its use. Therefore, we revised the prior inventory to create a new one, the CLIH. This study describes results from psychometric assessment of the CLIH in the context of a simulated handoff performed by medical students. To provide evidence in support of the validity of the scores from this measure, the study examined the factor structure and determined whether the CLIH scores vary, as predicted by CLT, with a measure of overall cognitive load, learner experience, case complexity, and performance.

Methods

Design

This is a psychometric study of the CLIH in which we utilized the unitary model of validity[31,32] to obtain evidence from several sources: content of the items (input from experts), response process (cognitive think aloud with residents), internal structure (factor analysis and internal consistency), and correlation with other variables. We did not collect evidence for consequential validity.

Development of the CLIH (content validity)

To guide item development, the authors focused on recently published conceptual work that identifies drivers of cognitive load types during handoffs[7] and also on emerging empirical work that has reported success in measuring cognitive load types with self-report instruments—two studies of college students learning classroom material[14,28] and one study of medical trainees performing colonoscopy.[29] The IL items from our initial study[30] were modified and several new items added to capture hypothesized drivers were as follows: the volume, complexity, and interactivity of the handoff information. We wrote new EL and GL items. EL included items focused not only on task design (e.g. clarity of the protocol), but also recently proposed dimensions such as the physical[33] and internal[30] environment. Following the recommendations of several studies,[12,30] concepts related to schema construction and metacognition (e.g. taking steps to clarify understanding) were adapted to further specify GL. Three CLT and

Table 1. Factor loadings[a] for each handoff simulation.

Item	Item content[b]	Case A		Case B		Case C	Case D		
		F1	F2	F1	F2	F1	F1	F2	F3
IL1	I found the volume of clinical information difficult to process.	0.99	0.40	0.92		0.81	0.86		
IL2	The patient problems were complex.	0.76		0.83		0.93	0.72		
IL3	My uncertainty about the diagnosis, prognosis or plan made it difficult to establish a clear picture of the current clinical situation.	0.63		0.84		0.87	0.58		0.47
IL4	My own knowledge gaps made it difficult for me to understand the patient problems.	0.57		0.79		0.93	0.59		
IL5	The potential for interactions between diagnoses and/or treatments added complexity.	0.64		0.55	0.53	0.76	0.59		
IL6	This patient was difficult to characterize with a one-line summary.	0.73		0.58		0.63	0.74		
IL7	I needed more time to establish an understanding of the clinical information.	0.79		0.87		0.85	0.68	0.41	
EL1	The protocol that I was expected to use for the sign-out (SBAR) was not clear to me.				0.95			0.68	
EL2	The terminology used by the intern was difficult to understand.			0.47		0.45		0.73	
EL3	I felt distracted by the environment (such as environmental noise and the layout of the room).								
EL4	I felt distracted by things on my mind unrelated to the sign-out.		0.56		0.40				
EL5	I felt distracted by worries about whether I was performing the handoff adequately.							0.57	
EL6	I was distracted by the intern's style of communicating (too fast/slow, too much/little detail, over- or under-confident).								
GL1	I invested substantial mental effort trying to connect my knowledge to the patient problems.	0.64		0.90		0.69	0.59		
GL2	I invested substantial effort in mentally organizing the patient information into a coherent clinical picture.	0.88		0.88		0.84	0.75		
GL3	I invested substantial mental effort in remembering the patient problems.	0.82		0.73		0.73	0.74		
GL4	I invested substantial mental effort in understanding the patient problems.	0.46		0.85		0.87	0.55		
GL5	I invested substantial mental effort in taking steps to clarify my understanding.	0.76		0.70		0.69	0.87		
GL6	I invested substantial mental effort in trying to follow the sign-out protocol.				0.45				

IL: intrinsic load; EL: extraneous load; GL: germane load; SBAR: situation, background, assessment, and recommendation.
[a]Factor loadings derived from principal axis factoring with promax rotation using robust weighted least square estimation.
[b]The instructions for each item were as follows: "Please rate your level of agreement with each of the following statements regarding this handoff." Participants indicated their level of agreement via a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree).

two handoffs experts iteratively reviewed drafts. To examine the response process of trainees, five residents at the lead author's institution explained in a group setting how they understood each item leading to revisions of several items. These steps led to a significantly revised and expanded inventory requiring collection of new evidence for validity. The resulting CLIH had 19 items total with 7 items for IL and 6 items each for EL and GL (Table 1). The instructions for each item were as follows: "Please rate your level of agreement with each of the following statements regarding this handoff." Participants indicated their level of agreement via a 5-point Likert scale (strongly disagree to strongly agree).

Participants and procedures

The data for this study were collected in the context of a separate study that examined predictors of information loss and distortion during simulated handoffs.[34] Study participants were second- and sixth-year students, recruited from a Dutch medical school. Risk and benefits were explained to each participant and written informed consent was obtained. After providing information about prior handoff experiences, each participant performed four simulated handoffs. Two were simple cases; two were complex cases. Simple cases had an established diagnosis with typical associated clinical findings, whereas complex cases contained unrelated findings partially consistent with multiple possible diagnoses. The order of the cases was randomly assigned to each participant. At the end of each handoff, study participants completed the CLIH, Paas' Cognitive Load Scale, and one global rating item for IL, EL, and GL, respectively. This study was approved by the Ethical Review Board of the Netherlands Association for Medical Education.

Relationship with other variables

We assessed the relationship of the CLIH with several variables. First, the total cognitive load score from the CLIH should

correlate positively with measures of overall cognitive load. To test this hypothesis, we adapted Paas' Cognitive Load Scale (Paas' Scale), a single item designed to measure overall cognitive load to read: "During the handoff I just finished, I invested …" followed by a 9-point scale (ranging from extremely low mental effort to extremely high mental effort). Second, the score of each load type should correlate with a global measure of each load type, respectively. We, therefore, included a single global item for IL, EL, and GL, respectively. We assumed that the overall perception of each kind of load will correlate with the corresponding score generated by the CLIH.

Finally, the CLIH should relate to other variables as predicted by CLT. For example, as a learner's knowledge increases, a given task becomes more routine and the IL and the total cognitive load for that task should decrease. We, therefore, examined the relationship between the CLIH and experience (level of training), task complexity (simple vs complex cases), and performance (proportion of information successfully transferred during the simulated handoff).

For the performance outcome, we used a measure of the proportion of information successfully transferred during each simulated handoff that had been calculated for a different study occurring during the same simulation. Details are described elsewhere.[34] In short, for each case signed-out by a subject, an overall index of information accuracy was calculated. Because cases differed in their total number of information elements, the raw scores were standardized.

Analysis

There were four different handoff cases that each participant completed, that is, four case results were nested within each participant. Given the nested structure of this data, we preferred to conduct a multilevel factor analysis. However, our limited sample size did not permit this approach. We, therefore, performed separate categorical exploratory factor analysis (EFA) on each of the four cases. We conducted principal axis factoring with promax rotation using robust weighted least square estimation in Mplus (Muthen and Muthen[35]). We included all 19 items. For model selection (number of factors extracted), we used several criteria to indicate acceptable fit: the eigenvalue must be greater than 1, a factor must have at least two items with loadings greater than 0.40, and the standardized root-mean-square residuals should be less than 0.08 (Hu and Bentler[36]). To derive the final model selection, we examined the consistency in patterns of factor loadings across the four simulations.

We created scores for the resulting factors by summing the items that composed of the factor. SPSS version 23.0 (IBM Corporation, Armonk, NY) was used for the additional analyses. Because each subject participated in all four conditions (two simple cases plus two complex cases), we used repeated-measures analysis of variance to assess the CLIH score by case complexity (within subjects) and level of training (between subjects).

Results

Participant characteristics

In total, 52 medical students participated. A typical participant was a female (N = 47, 85%) sixth-year student (n = 29, 56%). Approximately half of the participants reported performing fewer than five handoffs as a sender and as a receiver. Most (N = 40, 78%) had not received any prior training in handoffs. Because level of training and reported number of prior handoffs covaried, we used only level of training as the experience variable in subsequent analyses.

Factor analyses

Each simulation generated a unique model (Table 1). The number of factors identified in each of the four models varied from two factors in cases A and B, one factor in case C, and three factors in case D (Table 1). The differences between the models were largely due to the inconsistencies in the distribution of EL items across the factors. Yet, all four models had a significant common feature: a single IL/GL factor that represented all of the IL items and five out of the six GL items. Moreover, internal consistency for the IL/GL factor was high (Cronbach's alpha was 0.92). Based on these results, we concluded that one factor with the overlapping indicators of IL and GL was stable and met our model-fit criteria.

Additional evidence for validity

Given the consistency of the one factor representing the IL and five of the six GL items across the four simulations, we created an IL/GL factor to explore additional evidence for validity. Table 2 summarizes the relationship of the single factor incorporating IL/GL with the other variables. For three of the variables (Paas' Scale score, IL/GL score, and performance score), each study participant had four scores (one for each completed case) and these scores had high internal consistency (Cronbach's alpha: 0.89 for the IL/GL score, 0.83 for Paas' Scale score, and 0.95 for the performance score). Therefore, for each variable, we used the means score across the four cases for each individual to assess correlations.

The mean IL/GL score correlated with the mean score of the Paas measures ($r = 0.83$, $p < 0.001$). In addition, the IL/GL score correlated negatively with handoff accuracy (Pearson's $r = -0.34$, $p < 0.001$). Finally, the IL/GL score correlated highly with both the Global IL item (Pearson's $r = 0.81$, $p < 0.001$) and the Global GL item (Pearson's $r = 0.83$, $p < 0.001$). No correlation was calculated for Global EL since we did not identify an EL factor. Between-subjects, repeated-measures regression analysis showed that IL/GL was lower for sixth-year students compared to second year students (beta $= -0.96$, $p < 0.001$) and simple versus complex cases (beta $= 0.74$, $p < 0.001$). There was no interaction between the two independent variables.

Table 2. Relationship of the IL/GL factor with other variables.

Variable	Hypothesized relationship with IL/GL	Result
Correlations (Pearson's r)		
Paas' Scale[1–9]	IL/GL increases with Paas' Scale	Pearson's[a] r = 0.83 (p < 0.001)
Performance[b]	IL/GL decreases as performance increases	Pearson's[a] r = −0.34 (p < 0.001)
Global IL item: "Overall I found the handoff challenging"	IL/GL increases as the Global IL item increases	Pearson's[a] r = 0.81(p < 0.001)
Global GL item: "Overall I invested substantial effort in learning during this handoff"	IL/GL increases as the Global GL item increases	Pearson's[a] r = 0.83 (p < 0.001)
Repeated-measures regression		
Between-subjects independent variable: year of training (second-year vs sixth-year students)	IL/GL decreases as level of training increases	Year of training beta, CI, p-value: −0.96, (−1.1 to −0.78), p < 0.001
Within subjects independent variable: case complexity (simple vs complex)	IL/GL increases with case complexity	Case complexity beta, CI, p-value: 0.74, (0.46 to 1.0), p < 0.001

CI: confidence interval; IL: intrinsic load; EL: extraneous load; GL: germane load.
N = 52 unless indicated.
[a]For three of the measures (Paas' Scale score, IL/GL score, and performance score), each study participant has four scores (one for each completed case). For each of these three measures, the scores had high internal consistency (see results). Therefore, we report here only the correlation between the mean scores of each individual.
[b]Performance: proportion of information accurately transmitted at handoff, N = 49.

Discussion

A measure that differentiates cognitive load types is necessary to identify the factors that most impact trainee learning and performance. The results of this study suggest that for handoffs, the CLIH measures a single dimension of cognitive load that is a combination of IL and GL. The score from this measure had high internal consistency and correlated as hypothesized with Paas' Scale, the global IL and GL items, level of training, case complexity, and performance. At the same time, the EL items did not perform as expected in the factor analyses. This represents our second failed attempt to measure EL during handoffs via subjective self-report.

The finding that IL and GL form one factor adds to the current debate among CLT and medical education researchers. CLT originally proposed two types of cognitive load: IL and EL.[8] In the 1990s, this framework was expanded to include a third factor, GL.[13] Yet, in recent years, some theorists have argued that GL does not constitute a separate type of load but rather falls within IL and represents the WM resources dedicated to processing IL.[15,37] Our results support this view that GL and IL are best understood as part of a single process, at least as perceived by students completing the CLIH. These conflicting results regarding the relationship of the GL items with IL items may highlight the challenge trainees face in assessing their own mental processes or the challenge in measuring a construct as complicated as handoffs.

While this instrument can be used to measure IL/GL during a handoff, it cannot be used to measure EL. The EL items did not form a single factor. There are several potential explanations. First, the sample size may have been too small to permit the detection of the underlying relationship between the EL items. Second, while our six items focused on the main drivers of EL (task design and organization)[38] and the physical environment[33], the construct of EL may nevertheless have been under-represented in our items. Other groups have reported difficulty measuring EL. In a recent mixed-methods study, Naismith et al.[25] present qualitative data suggesting that Paas' Scale, the NASA-TLX Scale, and their own Cognitive Load Component Measure do not adequately capture EL. Third, despite the pre-testing with five residents, the construction of the items themselves may not be sufficiently clear, and, as a result, the items may not be understood in a consistent manner across study participants or may be interpreted in a way that is not consistent with the construct. Finally, the context of the simulated handoff may be a significant factor. The physical environment was controlled to minimize EL from the physical environment (interruptions, noise, and space) and the design of the task (clear instructions, all needed information in a single place). In retrospect, this likely represents the most important reason why the EL items did not form a factor. Therefore, instead of abandoning efforts to assess EL in handoffs, we advocate exploring this construct in settings where EL factors are not intentionally minimized.

Additional limitations of the study included the size of the nested sample (208 observations nested within 52 subjects). Multilevel factor analysis would have been preferable, but the sample size was too small for this approach. Another limitation was that participants were recruited from a single institution. Strengths included two different levels of learners, varying case complexity, and a relatively robust measure of performance.

In summary, this study provides several sources of validity evidence for the IL/GL score generated by the CLIH. The EL

items did not perform well. Yet, differentiating EL from IL/GL is essential if we are to use CLT to improve the instructional and clinical environments. Therefore, the EL items need to be redrafted with a more systematic assessment of the response process. In addition, the next version should be tested in either a simulated environment that intentionally introduces and varies EL or, even better, an authentic clinical setting. A measure that can differentiate between IL/GL and EL would allow handoffs researchers to determine the relative contribution of EL and IL/GL to handoff errors. This would help prioritize efforts to improve patient safety during handoffs. Current handoff protocols focus on reducing EL rather than managing IL.[39] But, we do not know how effective these practices are in reducing EL if we cannot measure it; nor do we know whether the emphasis on EL is warranted. Moreover, the ability to measure load types would enable us to better understand the cognitive mechanisms and effectiveness of current and future handoff interventions and develop a bundle that effectively manages all cognitive load types.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval

Ethical approval for this study was obtained from the Ethical Review Board of the Netherlands Association for Medical Education. NERB Dosier no. 450.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Informed consent

Written informed consent was obtained from all subjects before the study.

References

1. Horwitz LI, Moin T, Krumholz HM, et al. Consequences of inadequate sign-out for patient care. *Arch Intern Med* 2008; 168(16): 1755–1760.
2. Riesenberg LA, Leitzsch J, Massucci JL, et al. Residents' and attending physicians' handoffs: a systematic review of the literature. *Acad Med* 2009; 84(12): 1775–1787.
3. Starmer AJ, O'Toole JK, Rosenbluth G, et al. Development, implementation, and dissemination of the I-PASS handoff curriculum: a multisite educational intervention to improve patient handoffs. *Acad Med* 2014; 89(6): 876–884.
4. Patterson ES, Roth EM, Woods DD, et al. Handoff strategies in settings with high consequences for failure: lessons for health care operations. *Int J Qual Health Care* 2004; 16(2): 125–132.
5. Wohlauer MV, Arora VM, Horwitz LI, et al. The patient handoff: a comprehensive curricular blueprint for resident education to improve continuity of care. *Acad Med* 2012; 87(4): 411–418.
6. Starmer AJ, Spector ND, Srivastava R, et al. Changes in medical errors after implementation of a handoff program. *N Engl J Med* 2014; 371(19): 1803–1812.
7. Young JQ, Ten Cate O, O'Sullivan PS, et al. Unpacking the complexity of patient handoffs through the lens of cognitive load theory. *Teach Learn Med* 2016; 28(1): 88–96.
8. Sweller J. Cognitive load during problem solving: effects on learning. *Cognitive Sci* 1988; 12(2): 257–285.
9. Sweller J and van Merrienboer JJG. Cognitive load theory and instructional design for medical e. In: Walsh K (ed.) *The Oxford textbook of medical education*. Oxford: Oxford University Press, 2013, pp. 74–85.
10. Baddeley A. Working memory: theories, models, and controversies. *Annu Rev Psychol* 2012; 63: 1–29.
11. Cowan N. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav Brain Sci* 2001; 24(1): 87–114; discussion 114–185.
12. Young JQ, Van Merrienboer J, Durning S, et al. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. *Med Teach* 2014; 36(5): 371–384.
13. Sweller J, van Merrienboer JJG and Paas FGWC. Cognitive architecture and instructional design. *Educ Psychol Rev* 1998; 10(3): 251–296.
14. Leppink J, Paas F, van Gog T, et al. Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learn Instr* 2014; 30: 32–42.
15. Sweller J, Ayres PL and Kalyuga S. *Cognitive load theory*. New York: Springer, 2011, p. xvi, p. 274.
16. Van Merriënboer JJG and Sweller J. Cognitive load theory and complex learning: recent developments and future directions. *Educ Psychol Rev* 2005; 17(2): 147–177.
17. De Leeuw KE and Mayer RE. A comparison of three measures of cognitive load: evidence for separable measures of intrinsic, extraneous, and germane load. *J Educ Psychol* 2008; 100(1): 223–234.
18. Paas FG. Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J Educ Psychol* 1992; 84(4): 429–434.
19. Cierniak G, Scheiter K and Gerjets P. Explaining the split-attention effect: is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Comput Hum Behav* 2009; 25(2): 315–324.
20. Haji FA, Rojas D, Childs R, et al. Measuring cognitive load: performance, mental effort and simulation task complexity. *Med Educ* 2015; 49(8): 815–827.
21. Naismith LM and Cavalcanti RB. Validity of cognitive load measures in simulation-based training: a systematic review. *Acad Med* 2015; 90(11 Suppl.): S24–S35.
22. Galy E, Cariou M and Mélan C. What is the relationship between mental workload factors and cognitive load types? *Int J Psychophysiol* 2012; 83(3): 269–275.
23. Paas F, Tuovinen JE, Tabbers H, et al. Cognitive load measurement as a means to advance cognitive load theory. *Educ Psychol* 2003; 38(1): 63–71.
24. Van Gog T and Paas F. Instructional efficiency: revisiting the original construct in educational research. *Educ Psychol* 2008; 43(1): 16–26.
25. Naismith LM, Cheung JJH, Ringsted C, et al. Limitations of subjective cognitive load measures in simulation-based procedural training. *Med Educ* 2015; 49(8): 805–814.

26. Hart SG and Staveland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock PA and Meshtaki N (eds) *Human mental workload*. Amsterdam: North-Holland, 1988, pp. 139–183.
27. Hart SG and Staveland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research: Ames Research Center, 1988, https://archive.org/details/nasa_techdoc_20000004342
28. Leppink J, Paas F, Van der Vleuten CP, et al. Development of an instrument for measuring different types of cognitive load. *Behav Res Methods* 2013; 45(4): 1058–1072.
29. Sewell JL, Boscardin CK, Young JQ, et al. Measuring cognitive load during procedural skills training with colonoscopy as an exemplar. *Med Educ* 2016; 50(6): 682–692.
30. Young JQ, Irby DM, Barilla-LaBarca ML, et al. Measuring cognitive load: mixed results from a handover simulation for medical students. *Perspect Med Educ* 2016; 5: 24–32.
31. Kane MT. Current concerns in validity theory. *J Educ Meas* 2001; 38(4): 319–342.
32. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003; 37(9): 830–837.
33. Choi H-H, Van Merriënboer JJG and Paas F. Effects of the physical environment on cognitive load and learning: towards a new model of cognitive load. *Educ Psychol Rev* 2014; 26(2): 225–244.
34. Young JQ, van Dijk SM, O'Sullivan PS, et al. Influence of learner knowledge and case complexity on handover accuracy and cognitive load: results from a simulation study. *Med Educ* 2016; 50(9): 969–978.
35. Muthen LK and Muthen BO. *Mplus user's guide*. 7th ed. Los Angeles, CA: Muthen & Muthen, 2012
36. Hu L and Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Modeling* 1999; 6(1): 1–55.
37. Kalyuga S. Cognitive load theory: how many types of load does it really need? *Educ Psychol Rev* 2011; 23(1): 1–19.
38. Sweller J and Chandler P. Why some material is difficult to learn. *Cognition Instruct* 1994; 12(3): 185–233.
39. Young JQ, Wachter RM, Ten Cate O, et al. Advancing the next generation of handover research and practice with cognitive load theory. *BMJ Qual Saf* 2016; 25(2): 66–70.