DR. JOHN Q YOUNG (Orcid ID : 0000-0003-2219-5657)

DR. JUSTIN L SEWELL (Orcid ID : 0000-0003-4049-2874)

DR. KAREN FRIEDMAN (Orcid ID : 0000-0003-1980-1839)

Article type : Original Research

Evidence for Validity for the Cognitive Load Inventory for Handoffs

John Q. Young, MD, MPP, PhD, Majnu John, MS, PhD, Krima Thakker, Karen Friedman, MD, Rebekah Sugarman, Justin L. Sewell, MD, MPH, PhD, Patricia S. O'Sullivan, EdD

Dr. Young is Professor of Psychiatry, Department of Psychiatry, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, and the Zucker Hillside Hospital at Northwell Health, Glen Oaks, NY.

Dr. John is statistician, Division of Research, Zucker Hillside Hospital at Northwell Health, Glen Oaks, NY.

Ms. Thakker is research coordinator, Division of Education and Training, Zucker Hillside Hospital at Northwell Health, Glen Oaks, NY.

Dr. Friedman is Associate Professor of Medicine, Department of Medicine, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY.

Ms. Sugarman is a medical student, University of Michigan School of Medicine, Ann Arbor, MI.

Dr. Sewell is Associate Professor, Department of Medicine, University of California, San Francisco School of Medicine, San Francisco, CA.

Dr. O'Sullivan is Professor, Department of Medicine, and Director of Research and Development in Medical Education, University of California, San Francisco School of Medicine, San Francisco, CA.

Corresponding Author:

John Q. Young, MD, MPP, PhD
Department of Psychiatry
The Zucker Hillside Hospital
Zucker School of Medicine
75-59 263rd Street
Glen Oaks, New York, 10543 USA
Jyoung9@northwell.edu

ABSTRACT

OBJECTIVE

Patient handoffs remain a significant patient safety challenge. Cognitive Load Theory (CLT) can be used to identify the cognitive mechanisms of handoff errors. The ability to measure cognitive load types during handoffs could drive the development of effective protocols and educational strategies. No such measure currently exists.

METHOD

The authors developed the Cognitive Load Inventory for Handoffs (CLIH) using a multi-step process, including expert interview to enhance content validity and talk-alouds to optimize response process validity. The final versions contained 28 items. From January to March, 2019, we administered a cross-sectional survey to 1,807 residents and fellows from a large health system in the United States. Participants completed the CLIH following a handoff. Exploratory factor analysis from one-third of respondents identified high performing items; confirmatory factor analysis with the remaining sample assessed model fit. Model fit was evaluated using the comparative fit index (>0.90), Tucker-Lewis Index (>0.80), standardized root mean square residual (<0.08), and root mean square of approximation (<0.08).

RESULTS

Participants included 693 trainees (38.4%) (231 in the exploratory study and 462 in the confirmatory study). Eleven items were removed during exploratory factor analysis. Confirmatory factor analysis of the 16 remaining items (5 intrinsic load, 7 extraneous load, and 4 germane load) supported a three-factor model and met criteria for good model fit: CFI was 0.93, RMSEA was 0.074, and SRMR was 0.07. The fit was comparable for gender and role. Intrinsic, extraneous and germane scales had high internal consistency. With one exception, scale scores associated, as hypothesized, with postgraduate level and clinical setting.

CONCLUSION

The results provide evidence for validity: content, response process, internal structure and association with other variables. This instrument can be used to determine the relative drivers of cognitive load during handoffs as well as the relative effectiveness of handoff instruction and protocols.

INTRODUCTION

Patient handoffs is the process by which the clinical information and responsibility for a patient or panel of patients is transferred from one clinician or team to another.[1] Handoffs may be triggered by any number of events, including the end of the clinician's shift or change in a patient's level of care.[2] Handoffs may occur frequently in any given day and are vulnerable to communication errors.[3,4] Communication errors include information loss and/or distortion that can lead to patient harm.[5-7] Efforts to improve patient safety during handoffs have generated best practices.[8,9] These practices facilitate information transfer via communication protocols that include structured face-to-face and written sign-out, teamwork, interactive questioning, and distraction-free settings.[10,11] In an effort to improve the competencies of physicians-in-training, medical schools and residency programs are implementing handoff curricula that teach these best practices.[8]

Despite these advances, handoffs remain a significant patient safety challenge, even in those studies reporting improvements.[11] Researchers have identified Cognitive Load Theory as a framework to explore the cognitive mechanisms of handoff errors and thereby develop evidence-based handoff processes for which educators can design specific interventions.[12,13] Originally developed by John Sweller in the context of studying how students problem solve[14], Cognitive Load Theory (CLT) focuses on the implications of limited working memory (WM) for learning.[15] While sensory and long-term memory are relatively infinite, WM is finite. In fact, WM can only actively process (i.e. organize, compare and contrast) two to four elements at any given moment.[16,17] When the cognitive load of a learning tasks such as a handoff exceeds the working memory capacity of the trainee, 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). This can lead to patient harm.

CLT envisions at least two, and possibly three, types of cognitive load (CL).[18] *Intrinsic load* arises from the information processing demands associated with performing the task itself. Both task complexity and learner expertise determine the intrinsic load imposed by any given handoff. *Extraneous load* occurs when learners use working memory resources to process information not essential to the task. Common examples include external distractions (e.g., noise in the environment) or suboptimal instructional design (e.g., unnecessarily having to search for information).[19] Internal distractions (e.g., worries about external or personal issues, competing demands, self-induced time pressure) may also contribute to extraneous load.[20,21] Germane load represents the information processing imposed by the learner's deliberate use of cognitive strategies to refine existing schemata and enhance storage in long-term memory.[22] Recent work by Sweller and others has suggested that germane load may best be understood as a component of intrinsic load rather than a separate type of load.[23-25] Yet, some empirical work has found evidence for germane load as a separate type.[26,27]

In order to identify the cognitive mechanisms of current handoff protocols and to develop new handoff strategies that modulate intrinsic, extraneous, and germane loads in the desired directions, we need measures that differentiate between the types of cognitive load impacting a learner during a handoff. To date, two studies have attempted to develop a measure of cognitive load types during handoffs.[26-28] The extraneous load items in both studies performed poorly and the findings related to germane load were contradictory. To build upon this prior work, we revised the Cognitive Load Inventory for Handoffs (CLIH) and then collected evidence for validity.

METHOD

Study design and ethics

This is a psychometric study of an instrument (CLIH) designed to measure the cognitive load experienced by trainees during handoff of a patient panel. Consistent with the work of Downing and Kane and current standards in educatoinal and psychological measurement, we used the unitary model of construct validity, with multiple sources of evidence, including: content, response process, internal structure, and relationship with other variables.[29-31]

The Institutional Review Board for Northwell Health reviewed and deemed the study protocol exempt status as an educational research study with minimal risk.

CLIH development

We followed recommendations in the literature for survey development.[32] We reviewed previously published literature on cognitive load theory and medical education[18,33], drivers of cognitive load during handoffs[12], instruments that successfully measured cognitive load during colonoscopy and classroom learning[26,34], and earlier versions of the CLIH that reported mixed and contradictory results.[27,28] Based on insights from these prior studies, we drafted items for each type of cognitive load. Each item was mapped to a CLT construct and handoff feature. (Supplementary Online Appendix 1) For example, an intrinsic load item asked about the difficulty in managing the amount of clinical information. This item was mapped to the CLT construct of number of information elements and to handoff features such as the number of patients, comorbidities per patient, and follow-up tasks. Similarly, another item asked about the complexity of the patient problems which was mapped to the CLT construct of element interactivity and the handoff features of uncertainty, interactions, and maturity of the evidence base for the disease. The first author then conducted individual interviews with 9 international experts in cognitive load theory (5) and handoffs (4). These experts provided feedback on each item with respect to the clarity and alignment with the intended CLT construct and handoff feature. Items were revised, deleted, and/or added after each interview. Adaptation of items from previously published work, development of additional items to ensure representativeness to the domains of CLT and handoffs, and expert review enhanced item quality and content evidence of validity.

We conducted several types of pilot studies and cognitive interviews with trainees in order to reduce construct-irrelevant variance and optimize the response process, i.e., the probability that the resident interpreted each item as intended by its authors. First, we (JQY and RS) performed interviews with two groups (5 chief residents in psychiatry and 12 third year residents in psychiatry). Then we conducted individual interviews with 7 internal medicine and

2 pediatric residents. All of these residents were identified by their program director and agreed to participate. During the group and individual interviews, each item was read and residents were asked to say aloud what they thought the item was asking. When JQY and/or RS perceived either confusion or a discrepancy between the intended and perceived meaning of an item, follow up questions were asked to identify the source and potential solutions. Items were revised after each interview. We stopped the cognitive interviews when two consecutive interviews surfaced no discrepancies.

All authors approved the final version, which included 28 items, each employing an eleven point scale (0 for strongly disagree to 10 for strongly agree). Ten items measured IL, twelve items measured EL, and six items measured GL. We intentionally included a larger number of items than we intended to ultimately retain in order to minimize construct under-representation, especially for EL which two prior studies had failed to measure.[23] We also developed three global rating items, one for each CL type, for internal validation. Supplementary Online Appendix 2 shows the 28 CLIH items and 3 global items.

Participants and procedures

We recruited residents and fellows from Northwell Health, a large, twenty four hospital health system in the New York City metropolitan area that sponsors 122 distinct Accreditation Council for Graduate Medical Education accredited residency and fellowship programs. The Office of Academic Affairs provided email addresses for the 1,823 residents and fellows active during the 2018-2019 academic year. Between January and March, 2019, each trainee received an email invitation from three study authors (JQY, KF, RS) with a link to the electronic survey hosted by RED-Cap, an academic software program that supports research surveys.[35] In addition to the CLIH and global items, respondents provided demographic data, including gender, year in training, specialty of training, service and setting in which the handoff occurred, reason for the handoff, and the number of hours since the handoff was completed. Non-respondents received weekly emails over seven weeks in order to increase response rate.[36] Reminders were sent at different times of day (0600, 0900, 1200, and 1700) to capture the transition between night-day and day-night shifts. We asked participants to complete the survey after a handoff. Invitees could participate only once and could enter a drawing for one of four $250 gift cards.[37]

Outcomes and Analysis

We obtained evidence for validity from several sources: content, response process, internal structure, and relationship with other variables.

Content evidence and response process. The content evidence for validity derived from the instrument development process described above, in which item development incorporated prior published research, systematic mapping to CLT and handoff constructs, iterative revisions based on input from relevant experts, and inclusion of more items than expected to be retained to minimize construct under-representation. The validity of the response process was enhanced by the multiple cognitive interviews which helped identify sources of confusion in the items and thereby reduce construct-irrelevant variance.

Internal Structure. Despite several prior studies, an instrument development process that included expert consultation, and the assumption of three factors, we did not know how many factors the items would form in practice. First, prior versions of the tool had not performed well, especially the items intended to measure EL and GL, leading to a one factor and two factor solution in two different studies.[27,28] Second, while we hypothesized three factors, strong arguments – both theoretic and empiric – have been made for both two factor and three factor models.[25] Finally, we added emotion items to measure extraneous load, which had not been used before in the measurement of cognitive load. We were not sure if the emotion items would map onto a single construct of EL or lead to two different EL factors. Due to these uncertainties about the factor structure of the CLIH, we pursued a two-step process. In step one, we conducted exploratory factor analysis (EFA) to assess performance of the individual items and to better understand the factor structure. In step two, we performed confirmatory factor analysis (CFA) for cross-validation of the factor structure, including evaluation of model fit.

We used a split-sample strategy.[38,39] One third of the total sample was randomly assigned for EFA analysis and there was no overlap between the samples used for EFA and CFA. Consistent with expert recommendations, data from the eleven point CLIH scale was treated as interval, and, therefore, parametric methods were used.[40,41] Ordinary Least Squares was employed for EFA; this approach produces solutions very similar to maximum likelihood even when the underlying matrices are badly behaved. All EFA analysis was conducted using the function 'fa' within the 'psych' package in R. In order to allow the items to more distinctly group into a factor, Varimax rotation, which maximizes the sum of the variance of the squared loadings, was applied to the minimum residual solution. EFA was done iteratively. To be included, a factor was required to have an eigenvalue greater than 1 and contain at least two items with loadings greater than 0.40. At each iteration an item was removed if the item was split across factors or if the corresponding factor loading was less than 0.40.

We performed CFA using PROC CALIS procedure in SAS. Model fit was evaluated using the comparative fit index (>0.90), Tucker-Lewis Index (>0.80), standardized root mean square residual (<0.08), and root mean square of approximation (<0.08).[42,43] CFA generates standardized path coefficients for each scale item, which represent the strength of association of each item with the factor and can be interpreted as factor loadings.

We conducted several additional tests to assess internal structure. First, given the conflicting data regarding two versus three factor models for cognitive load, we compared fit for both models to determine if GL was a separate source of variance. Second, to assess internal consistency, we examined Cronbach's alpha. Finally, in order to assess the validity and generalizability across various sub-populations, multi-group confirmatory factor analysis was utilized. We divided the sample by gender and by role (patient handoff information sender versus patient handoff information receiver) and conducted separate sub-group CFA Measurement invariance across groups was assessed by examining the invariance of patterns of factor loadings. The model was considered to be invariant across the groups if the difference in chi-square values between the unconstrained model and the weight constrained model was above the 5% significance level.

Relationship to Other Variables. We used two methods to assess relationships to other variables. First, we used Pearson's r to examine the correlation between the global rating items and total IL, EL, and GL. Second, we used univariate regression to analyze how each cognitive load type varied with level of training and with clinical setting. Because we hypothesized the most significant difference to be between interns and all others, we dichotomized the respondents accordingly. Similarly, we dichotomized clinical setting into ICU versus other setting because we hypothesized patient complexity to be higher in the ICU. We expected IL, EL, and GL to decrease for more advanced trainees and only IL to increase for the ICU.

RESULTS

Of the 1823 trainees invited to participate, 16 had undeliverable email addresses, resulting in a pool of 1807 potential participants. We received 693 responses (38.4%), representing all training programs in the health system. 231 were randomly assigned to the EFA and 462 to the CFA. Table 1 summarizes the characteristics of the respondents. The majority of the respondents were in their first three years of residency (78%) with males and females equally represented. Most handoffs were at end-of-shift (79%) and occurred in the non-ICU inpatient setting (67%). Participants came from all specialties, with the majority representing non-surgical disciplines (77%). The average number of patients per handoff was 10.3 but with a large standard deviation (10.5) indicating substantial diversity across settings. Overall, with only a few exceptions, participant characteristics were similar in the EFA and CFA groups. Statistical tests indicated significance in the number of handoffs occurring within a surgical service (greater proportion in the EFA sample) and in the number of handoffs occurring by a trainee in a non-surgical residency program (greater proportion in the CFA sample). About 27% of the respondents responded to reminders/invitations four through seven (roughly four to seven weeks after the initial invitation). Mean IL, EL, and GL did not differ between the two groups.

Internal Structure

Exploratory Factor Analysis. At the outset, two items were removed (EL2 and EL3) because they appeared to be not relevant to the sender role which represented 80% of the participants. In addition, three EL items (EL1, EL9, EL12), five IL items (Il1, IL5, IL6, IL9, IL10), and 2 GL items (GL5, GL6) were removed sequentially because they performed poorly (i.e., factor loading less than 0.40 and/or split across factors). The final model had 16 items (5 IL, 7 EL, and 4 GL) and produced a three factor (eigenvalues exceeding 1) model explaining 52% of the total variance. (Supplementary Online Appendix 2) Item loadings were high (0.52 to 0.90 for IL, 0.40 to 0.75 for EL, and 0.50 to 0.86 for GL) with only one cross-loading higher than 0.3.

Confirmatory Factor Analysis. Fifty one of the 462 surveys assigned to the CFA were incomplete and excluded from the analysis. All factor loadings were well above 0.50 and were statistically significant. (Table 2) Overall means (SD) were 4.76 (2.06) for IL, 2.65 (1.88) for EL, and 3.45 (2.29) for GL. Modification indices identified inter-correlation between two pairs of items (IL3/IL4 and EL7/EL8). We allowed these two pairs to correlate. In the modified three-factor model, the goodness-of-fit parameters were all favorable and exceeded our pre-determined thresholds (Table 3). The correlations between each factor were moderate: $r_{il, el}$= 0.40 ; $r_{il,gl}$= 0.52; and $r_{el,gl}$= 0.68. (Table 2). Figure 1 depicts the path diagram for the measurement model and highlights the factor structure, factor loadings, and correlations between the factors.

Additional internal structure evidence. The goodness-of-fit parameters were superior for the three-factor model compared to the two-factor model suggesting that GL was a separate source of variance and not nested within IL. (Table 3) Internal consistency of each factor was high; Cronbach's alpha was 0.85 for IL, 0.87 for EL, and 0.91 for GL. Finally, there was no difference between the unconstrained models and weight-constrained model for gender (Chi-square difference = 9.05, df = 13, p-value = 0.77) and role (sender versus receiver; Chi-square difference = 17.93, df = 13, p-value = 0.16). The factor structure was stable across these sub-groups.

Relationship with Other Variables

Pearson's correlations between the IL, EL, and GL scores and their respective global rating items (Table 2) were moderately strong for IL (0.51) and El (0.75) but small for GL (0.22). As predicted by CLT, more advanced trainees had lower EL and GL compared to interns. However, IL for more advanced trainees was not different. The relationship between cognitive load types and clinical setting confirmed the apriori hypotheses. IL was significantly higher in the ICU compared to other settings while there were no differences for El and GL. (Table 4)

DISCUSSION

In this study, we developed and tested an instrument designed to measure cognitive load experienced by trainees during patient handoffs. The item development process included a number of features that support validity in the content evidence and response process, including iterative revisions based on expert input, systematic mapping to CLT and handoff constructs, and multiple cognitive interviews with trainees. The results marshal strong evidence supporting the internal structure of the CLIH. EFA produced a three-factor structure with 16 high performing items. CFA confirmed the superiority of a three-factor model compared to a two-factor model with excellent fit indices. In addition, internal consistency was high and the factor structure did not differ for female versus male respondents. Finally, with a few exceptions to be discussed below, the mean IL, El, and GL scores varied as predicted with other variables. There was a particularly robust finding for clinical setting in which GL was significantly lower for more advanced trainees. Taken together, these findings are very encouraging and support the ability of the CLIH to measure IL, EL, and GL during handoffs. With such a measure educators can study tailored handoff interventions.

While the overall findings provide strong support for the CLIH, there were two specific results that were not expected. First, IL did not decrease for more advanced learners. In looking at the mean IL by each level of training, it appears that IL decreases modestly from intern year until the end of residency but then increases during the first year of fellowship. This finding likely reflects how the transition from senior resident to first year fellow puts the trainee in a new role and setting where their skills are relatively less developed. However, when we examined, in secondary analysis, how IL changes during residency only, the observed decrease was not statistically significant. This surprised us. One possible explanation is that as trainees advance during residency, their increased expertise is matched with more challenging roles that keeps IL more or less constant. The second inconsistent finding relates to the low correlation between the global rating item for GL and mean GL. In retrospect, we think this may relate to poor construction of the GL global rating item. The focus on 'activities to better understand the patient problems' may not adequately differentiate between IL and GL.

This study has several methodological strengths. Although CLT is considered highly applicable to patient handoffs, efforts to measure cognitive load have encountered difficulty. Two prior studies successfully measured IL but failed to form a stable factor structure for EL in either study and reported inconsistent findings regarding whether GL was best understood as a separate factor or nested within IL.[26-28] This study's success in measuring the three types of CL may be due to three key differences from the prior studies. First, item development included systematic mapping to CLT and handoff constructs and iterative revisions based on input from many experts. Second, in this study, we engaged in extensive evaluation of the response-process through numerous think-alouds, both with small groups and individuals. This process enabled us to identify and address multiple instances where the wording of the item was either confusing to trainees or interpreted in a manner different from what was intended. Finally, unlike the prior two studies which measured cognitive load in the context of simulated handoffs, this study asked trainees to complete the instrument after actual handoffs. This may have especially influenced the performance of the extraneous load items, since both prior studies on handoffs occurred in the context of simulations that intentionally removed distractions and other sources of extraneous load.[27,28] These methodological strengths may be helpful to the future development of medical education instruments in general and CLT inventories in particular.

This study also has important findings for CLT. Based on both empirical and theoretical research, it was plausible that either a four-factor model (where the internal distraction items form a separate factor in addition to IL, EL, and GL), three-factor model (including IL, EL, and GL) or a two-factor model (including only IL and EL) would provide the best fit for the data.[20,25,44] The superiority of the three factor solution has two significant implications. First, the internal distraction and interpersonal friction items loaded onto EL and did not form a separate factor. This opens up an entirely new dimension for investigating EL. To date, EL has been understood as mostly related to task design and more recently the environment (e.g., interruptions).[19,25] These results suggest that internal distraction (e.g., worries or self-consciousness and interpersonal friction (e.g., annoyance with the style of another person) contribute to EL. Future research should examine to what extent these factors influence learning and performance.

Second, researchers currently debate whether to conceptualize germane load as a third type of cognitive load distinct from intrinsic and extraneous load or as a subset of intrinsic load.[20,45-47] Many of the CLT researchers are now advocating for a two-factor model that understands GL as a component of IL.[25] In contrast, this study's results were most consistent with a three factor model. Interestingly, the only other instrument developed specifically for a medical education context is one that measures cognitive load during colonoscopy.[26] That measure also demonstrated three factors. While these results challenge the two-factor proponents, it is important to note that the third factor in both this and the colonoscopy study may not represent GL but another construct. GL, by definition, enhances learning. Examples of means to promote germane load include instructional design (e.g., interleaved practice compared to blocked practice) or prompting generative processes (e.g., self-explaining, or elaborating) [48]. Future studies can address this question by evaluating whether learning improves with instructional techniques that impact GL but not IL. Such studies must be careful to differentiate performance from learning and would be best to utilize outcomes that measure the impact of a technique several weeks later in a different context.

Finally, this study has implications for handoff research and educational practice. The CLIH can help researchers and educators identify strategies that improve learning and reduce errors during handoffs. With a measure that can differentiate cognitive load types, future studies will be able to identify to what extent a given handoff intervention effects each type of cognitive load. For example, to what extent does training monitoring one's understanding of the patients being discussed lead to higher GL? Do mindfulness techniques or deep breathing lead to reduced EL during handoffs? Does titrating patient complexity of a handoff panel to a resident's experience lead to less errors and/or improved learning? A tool like the CLIH allows practitioners to determine whether a given intervention or bundle of interventions influences IL, EL, and/or GL in the desired directions. Moreover, learners could complete the tool after handoff to help them identify by themselves or with the aid of a coach what was difficulty and how they might improve in their management of IL, EL, and GL.

The study has several limitations. The study's response rate is less than 40%. We do not know whether non-responders were different from responders. In addition, the study occurs in a single health system. However, this single health system is diverse and participants in the study came from multiple specialties and hospitals. The participants were only residents and fellows so we do not know how this instrument functions with students or faculty. Future studies should evaluate the stability of the factor structure in other populations and settings. Also, the CLIH is based on learner recall after the fact. More than a third of the participants completed the survey more than 24 hours after handoff. This is a significant length of time and introduces recall biases. For example, if transient but significant events such as distractions are poorly remembered, then the answers may under-report the impact of such factors. However, we are reassured by the results of a post-hoc analysis in which we performed sub-group CFA for time between handoff and completion of the CLIH (24 hours or less versus more than 24 hours). There was no difference in the factor structure between the two groups (Chi-square difference = 9.058, df = 13, p-value = 0.769) .

In conclusion, the CLIH shows evidence of measuring cognitive load types (IL, EL, and GL) during patient handoffs within a large sample of trainees from multiple specialties and hospitals. Improving handoff instruction requires strategies that reduce EL and optimize IL and GL. The CLIH should support such future efforts. The methodology used for the development of the CLIH, especially the close attention to response process, may help to improve the development of similar instruments in the future.

Acknowledgments: None

Funding: None

Ethical Approval: Ethical approval has been obtained from the Institutional Review Board for ZZZ (10/12/2018, IRB #: 18-0192)

Disclaimer: None

Previous Presentation: None

REFERENCES

1. 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.

2. Vidyarthi AR, Arora V, Schnipper JL, Wall SD, Wachter RM. Managing discontinuity in academic medical centers: strategies for a safe and effective resident sign-out. *J Hosp Med.* 2006;1(4):257-266.

3. Arora V, Johnson J, Lovinger D, Humphrey HJ, Meltzer DO. Communication failures in patient sign-out and suggestions for improvement: a critical incident analysis. *Qual Saf Health Care.* 2005;14(6):401-407.

4. Arora VM, Johnson JK, Meltzer DO, Humphrey HJ. A theoretical framework and competency-based approach to improving handoffs. *Qual Saf Health Care.* 2008;17(1):11-14.

5. Horwitz LI, Moin T, Krumholz HM, Wang L, Bradley EH. Consequences of inadequate sign-out for patient care. *Arch Intern Med.* 2008;168(16):1755-1760.

6. Gandhi TK, Kachalia A, Thomas EJ, et al. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. *Ann Intern Med.* 2006;145(7):488-496.

7. Young JQ, Eisendrath SJ. Enhancing patient safety and resident education during the academic year-end transfer of outpatients: lessons from the suicide of a psychiatric patient. *Acad Psychiatry.* 2011;35(1):54-57.

8. 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.

9. Patterson ES, Roth EM, Woods DD, Chow R, Gomes JO. Handoff strategies in settings with high consequences for failure: lessons for health care operations. *Int J Qual Health Care.* 2004;16(2):125-132.

10. 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.

11. 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.

12. Young JQ, Ten Cate O, O'Sullivan PS, Irby DM. Unpacking the Complexity of Patient Handoffs Through the Lens of Cognitive Load Theory. *Teach Learn Med.* 2016;28(1):88-96.

13. Young JQ, Wachter RM, Ten Cate O, O'Sullivan PS, Irby DM. Advancing the next generation of handover research and practice with cognitive load theory. *BMJ quality & safety.* 2016;25(2):66-70.

14. Sweller J. Cognitive load during problem solving: Effects on learning. *Cogn Sci.* 1988;12(2):257-285.

15. Sweller J, van Merrienboer JJG. Cognitive Load Theory and Instructional Design for Medical Education. In: Walsh K, ed. *The Oxford Textbook of Medical Education.* Oxford, UK: Oxford University Press; 2013:74-85.

16. Baddeley A. Working memory: theories, models, and controversies. *Annu Rev Psychol.* 2012;63:1-29.

17. 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.

18. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. *Med Teach.* 2014;36(5):371-384.

19. Choi H-H, van Merriënboer JJG, 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.

20. Young JQ, Sewell JL. Applying cognitive load theory to medical education: construct and measurement challenges. *Perspectives on medical education.* 2015;4(3):107-109.

21. Feldon DF. Cognitive load and classroom teaching: The double-edged sword of automaticity. *Educ Psych.* 2007;42(3):123-137.

22. Sweller J, van Merrienboer JJG, Paas FGWC. Cognitive architecture and instructional design. *Educ Psychol Rev.* 1998;10(3):251-296.

23. Leppink J, Paas F, van Gog T, van der Vleuten CPM, van Merrienboer JJG. Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction.* 2014;30:32-42.

24. Sweller J, Ayres PL, Kalyuga S. *Cognitive load theory.* New York: Springer; 2011.

25. Sweller J, van Merriënboer JJG, Paas F. Cognitive Architecture and Instructional Design: 20 Years Later. *Educ Psychol Rev.* 2019;31(2):261-292.

26. Sewell JL, Boscardin CK, Young JQ, ten Cate O, O'Sullivan PS. Measuring cognitive load during procedural skills training with colonoscopy as an exemplar. *Med Educ.* 2016;50(6):682-692.

27. Young JQ, Irby DM, Barilla-LaBarca ML, Ten Cate O, O'Sullivan PS. Measuring cognitive load: mixed results from a handover simulation for medical students. *Perspectives on medical education.* 2016;5(1):24-32.

28. Young JQ, Boscardin CK, van Dijk SM, et al. Performance of a cognitive load inventory during simulated handoffs: Evidence for validity. *SAGE open medicine.* 2016;4:2050312116682254.

29. Kane MT. Current concerns in validity theory. *Journal of Educational Measurement.* 2001;38(4):319-342.

30. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ.* 2003;37(9):830-837.

31. American Educational Research A, American Psychological A, National Council on Measurement in E, Joint Committee on Standards for E, Psychological T. *Standards for educational and psychological testing.* 2014.

32. Artino AR, Jr., La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No. 87. *Med Teach.* 2014;36(6):463-474.

33. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: Design principles and strategies. *Med Educ.* 2010;44(1):85-93.

34. Leppink J, Paas F, Van der Vleuten CP, Van Gog T, Van Merrienboer JJ. Development of an instrument for measuring different types of cognitive load. *Behav Res Methods.* 2013;45(4):1058-1072.

35. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics.* 2009;42(2):377-381.

36. Dillman DA, Smyth JD, Christian LM. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* Wiley Publishing; 2014.

37. Stovel RG, Ginsburg S, Stroud L, Cavalcanti RB, Devine LA. Incentives for recruiting trainee participants in medical education research. *Med Teach.* 2018;40(2):181-187.

38. Kyriazos T. Applied Psychometrics: The 3-Faced Construct Validation Methods, a Routine for Evaluating a Factor Structure. *Psychology.* 2018;9:2044-2072.

39. Woods CM, Edwards MC. 6 - Factor Analysis and Related Methods. In: Rao CR, Miller JP, Rao DC, eds. *Essential Statistical Methods for Medical Statistics.* Boston: North-Holland; 2011:174-201.

40. Cook DA, Beckman TJ. Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX. *Advances in health sciences education : theory and practice.* 2009;14(5):655-664.

41. Sullivan GM, Artino AR, Jr. Analyzing and interpreting data from likert-type scales. *J Grad Med Educ.* 2013;5(4):541-542.

42. Hooper D, Coughlan J, Mullen M. Structural equation modelling: Guidelines for determining model fit. *Articles.* 2008:2.

43. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling.* 1999;6(1):1-55.

44. Sewell JL, Maggio LA, Ten Cate O, van Gog T, Young JQ, O'Sullivan PS. Cognitive load theory for training health professionals in the workplace: A BEME review of studies among diverse professions: BEME Guide No. 53. *Med Teach.* 2019;41(3):256-270.

45. Leppink J, van den Heuvel A. The evolution of cognitive load theory and its application to medical education. *Perspectives on medical education.* 2015;4(3):119-127.

46. van Merrienboer JJ, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Medical education.* 2010;44(1):85-93.

47. Kalyuga S. Cognitive load theory: How many types of load does it really need? *Educational Psychology Review.* 2011;23(1):1-19.

48. Fiorella L, Mayer RE. Eight Ways to Promote Generative Learning. *Educational Psychology Review.* 2016;28(4):717-741.

Table 1: Characteristics of the participants

Characteristic		Total (N=693)		EFA (N=231)		CFA (N=462)		P-Vaue for
		n	%	n	%	n	%	EFA v CFA[i]
Year of training	PGY-1	215	31.02%	83	35.93%	132	28.57%	0.06
	PGY-2	180	25.97%	50	21.65%	130	28.14%	0.08
	PGY-3	144	20.78%	51	22.08%	93	20.13%	0.63
	PGY-4 Residents	50	7.22%	19	8.23%	40	8.66%	0.95
	PGY-4 Fellows	29	4.18%	6	2.60%	14	3.03%	0.93
	PGY-5 and higher	74	10.68%	22	9.52%	52	11.26%	0.57
	Missing	1	0.14%	0	0.00%	1	0.22%	1.0
Gender	Male	344	49.64%	104	45.02%	240	51.95%	0.10
	Female	343	49.49%	124	53.68%	219	47.40%	0.15
	Other	2	0.29%	2	0.87%	0	0.00%	0.21
	Prefer not to answer	3	0.43%	1	0.43%	2	0.43%	1.0
	Missing	1	0.14%	0	0.00%	1	0.22%	1.0

Table 1: Characteristics of the participants								
Characteristic		Total (N=693)		EFA (N=231)		CFA (N=462)		P-Vaue for
		n	%	n	%	n	%	EFA v CFA[i]
Clinical setting in which the handoff occurred	Inpatient ICU	90	12.99%	26	11.26%	64	13.85%	0.41
	Inpatient non-ICU	463	66.67%	160	69.26%	303	65.37%	0.35
	Emergency department	67	9.67%	21	9.09%	46	9.96%	0.83
	Ambulatory	28	4.04%	9	3.90%	19	4.11%	1.0
	Peri-operative setting	24	3.46%	9	3.90%	15	3.25%	0.82
	Other[a]	13	1.88%	3	1.30%	10	2.16%	0.62
	Missing	9	1.30%	3	1.30%	6	1.30%	1.0
Reason for the handoff	End of shift	550	79.37%	182	78.79%	368	79.65%	0.86
	Transfer to a different team within the same setting[b]	35	5.05%	12	5.19%	23	4.98%	1.0
	Transfer to a different setting[c]	36	5.19%	14	6.06%	22	4.76%	0.59
	End of rotation	61	8.80%	19	8.23%	42	9.09%	0.81
	Other[d]	2	0.29%	1	0.43%	1	0.22%	1.0
	Missing	9	1.30%	3	1.30%	6	1.30%	1.0
Number of hours since completion of the handoff	0 to 24 hours	366	52.81%	119	51.52%	247	53.46%	0.60
	24 hours to 5 days	119	17.17%	49	21.21%	70	15.15%	0.07
	More than 5 days	204	29.44%	63	27.27%	141	30.52%	0.39
	Missing	4	0.58%	0	0.00%	4	0.87%	0.38
Specialty of the service in which the handoff occurred	Surgical[e]	148	21.36%	60	25.97%	88	19.05%	0.04
	Non-Surgical[f]	533	76.91%	168	72.73%	365	79.00%	0.06
	Other	4	0.58%	1	0.43%	3	0.65%	0.54
	Missing	8	1.15%	2	0.87%	6	1.30%	0.90
Specialty of the trainee	Surgical[g]	146	21.07%	59	25.54%	87	18.83%	0.05

Table 1: Characteristics of the participants								
Characteristic		Total (N=693)		EFA (N=231)		CFA (N=462)		P-Vaue for
		n	%	n	%	n	%	EFA v CFA[i]
	Non-Surgical[h]	527	76.05%	163	70.56%	364	78.79%	0.02
	Other (Transitional Year)	17	2.45%	8	3.46%	9	1.95%	0.34
	Missing	3	0.43%	1	0.43%	2	0.43%	1.0
Role in Handoff	Sender	559	80.66%	178	77.06%	381	82.47%	0.10
	Receiver	125	18.04%	50	21.65%	75	16.23%	0.10
	Missing	9	1.30%	3	1.30%	6	1.30%	1.0
Number of patients per handoff		Mean (SD): 10.38 (10.56)		Mean (SD): 10.66 (11.06)		Mean (SD): 10.25 (10.31)		0.64

a. Call room, Lecture, Phone, Resident Quarters, Office, Between Shifts

b. For example, transfer from surgery to medicine.

c. For example, transfer from inpatient to outpatient.

d. Afternoon rounding, Communication between team members

e. Anesthesiology, General Surgery, Neurosurgery, Obstetrics and Gynecology, Oral Surgery, Orthopedics, Urology, Vascular Surgery

f. Cardiology, Critical Care, Dental, Dermatology, Emergency Medicine, Endocrinology, Ear/Nose/Throat, Family Medicine, Gastroenterology, Hematology/Oncology, Internal Medicine, Nephrology, Neurology, Neonatal Intensive Care Unit, Ophthalmology, Pediatrics, Pediatric Infectious Disease, Palliative Care, Physical Medicine and Rehabilitation, Podiatry, Pulmonary Medicine, Psychiatry, Radiology, Radiation Oncology, Rheumatology, Surgery Intensive Care Unit

g. Anesthesiology, General Surgery, Neurological Surgery, Obstetrics and Gynecology, Oral Surgery, Orthopedic Surgery, Plastic Surgery, Thoracic Surgery, Urology, Vascular Surgery

h. Dermatology, Emergency Medicine, Family Medicine, Internal Medicine, Neurology, Neuroradiology, Ophthalmology, Oral Pathology, Pathology, Pediatrics, Pediatric Dental Medicine, Physical Medicine and Rehabilitation, Podiatry, Psychiatry, Radiology, Radiation Oncology

i. P-Values are for chi square tests for proportions and t tests for means

Table 2 – Confirmatory Factor Analysis Results

Item #	Item	Mean (SD)	Factor Loading	Standard Error	P-value
	INTRINSIC LOAD: Please rate your agreement with the following statements regarding the handoff you have completed:				
IL 2	The patient problems were complex	5.34 (2.45)	0.62	0.03	<0.0001
IL 3[a]	The handoff included significant clinical decision(s) that needed to be made	4.99 (2.65)	0.49	0.04	<0.0001
IL 4[a]	The handoff included significant diagnostic and/or treatment uncertainty	4.23 (2.60)	0.53	0.04	<0.0001
IL 7	I had to consider multiple or complex interactions between diseases	4.54 (2.71)	0.92	0.01	<0.0001
IL 8	I had to consider multiple or complex interactions between treatments	4.50 (2.62)	0.93	0.01	<0.0001
Global IL	Overall, I found the patient problems difficult to understand.	-	-	-	-
	Overall Mean for IL[b,c]:	**4.76 (2.06)**			
	EXTRANEOUS LOAD: Please rate your agreement with the following statements regarding the handoff. These statements are about the environment and your mindset during the handoff:				
EL 4	The other clinician used jargon out of context	2.37 (2.35)	0.74	0.03	<0.0001
EL 5	I was distracted by the other clinician’s attitude	1.96 (2.20)	0.84	0.02	<0.0001
EL 6	I was self-conscious due to who was present	2.41 (2.60)	0.73	0.03	<0.0001
EL 7[a]	I was frequently interrupted (e.g., pages, phone calls, people, etc...)	3.37 (2.79)	0.56	0.04	<0.0001
EL 8[a]	Noise made it difficult to concentrate	3.00 (2.68)	0.66	0.03	<0.0001
EL 10	During the handoff, important information was not easily available when I needed it	2.56 (2.39)	0.77	0.02	<0.0001
EL 11	I was thinking about things unrelated to the sign-out	2.89 (2.55)	0.60	0.04	<0.0001
Global EL	Overall, I found it difficult to focus my attention on the handoff.	-	-	-	-
	Overall Mean for EL[b,c]:	**2.65 (1.88)**			

Table 2 – Confirmatory Factor Analysis Results

Item #	Item	Mean (SD)	Factor Loading	Standard Error	P-value
	GERMANE LOAD: Please rate your agreement with the following statements regarding your mental effort during the handoff you have completed:				
GL 1	I had to work hard to connect my own medical knowledge to the patient problems	3.18 (2.56)	0.88	0.01	<0.0001
GL 2	I had to work hard to organize the patient information into a coherent clinical picture	3.49 (2.61)	0.87	0.01	<0.0001
GL 3	During the sign-out, I had to work hard to concentrate on how well I understood the information	3.11 (2.53)	0.92	0.01	<0.0001
GL 4	I had to take steps to clarify points of confusion	3.95 (2.65)	0.71	0.03	<0.0001
Global GL	Overall, I invested mental effort in activities that helped me better understand the patient problems	-	-	-	-
	Overall Mean for GL[b,c,d]:	**3.45 (2.29)**			

Abbreviations: IL=Intrinsic Load; EL=Extraneous Load; GL=Germane Load

a. *Two pairs of items (IL3/IL4 and EL7/EL8) were allowed to correlate.*

b. Overall mean = sum of the items answered divided by the number of items answered. 11 point scale (strongly disagree to strongly agree).

c. Cronbach's alpha: Intrinsic Load = 0.85, Extraneous Load = 0.87, Germane Load = 0.91

d. The correlation between scales are as follows: $r_{il, el}$= 0.40 ; $r_{il,gl}$= 0.52; and $r_{el,gl}$= 0.68

Table 3: Measures of fit for two-factor and three-factor models

Model	χ^2, d.f., p-value, Normed χ^2	CFI[†]	Tucker-Lewis Index [ϕ]	RMSEA[‡] (95% CI)	SRMR[§]
Two-factor model	χ^2 = 1154.7, d.f. = 103, $p < 0.0001$, Normed χ^2 = 11.2	0.74	0.70	0.158 (0.150, 0.166)	0.1026
Two-factor model (modified)	χ^2 = 907.3, d.f. = 101, $p < 0.0001$, Normed χ^2 = 5.32	0.80	0.76	0.139 (0.1313, 0.1479)	0.0993
Three-factor model	χ^2 = 537.9, d.f. = 101, $p < 0.0001$, Normed χ^2 = 5.32	0.89	0.87	0.103 (0.0943, 0.1113)	0.0754
Three-factor model (modified)	χ^2 = 322.363, d.f. = 99, $p < 0.0001$, Normed χ^2 = 3.26	0.95	0.93	0.074 (0.065, 0.083)	0.0735

CFI = comparative fit index; CI = confidence interval; d.f. = degrees of freedom; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

[*] A non-significant ($p > 0.05$) χ^2 suggests the model is an adequate representation of the data. However, with large sample size (> 200), χ^2 is almost always significant, making the χ^2 fit index inappropriate for larger sample size data such as ours. Given the large sample size, the relative (normed) chi-square is recommended. This value equals the χ^2 index divided by the degrees of freedom. The criterion for acceptance is recommended as less than 5.

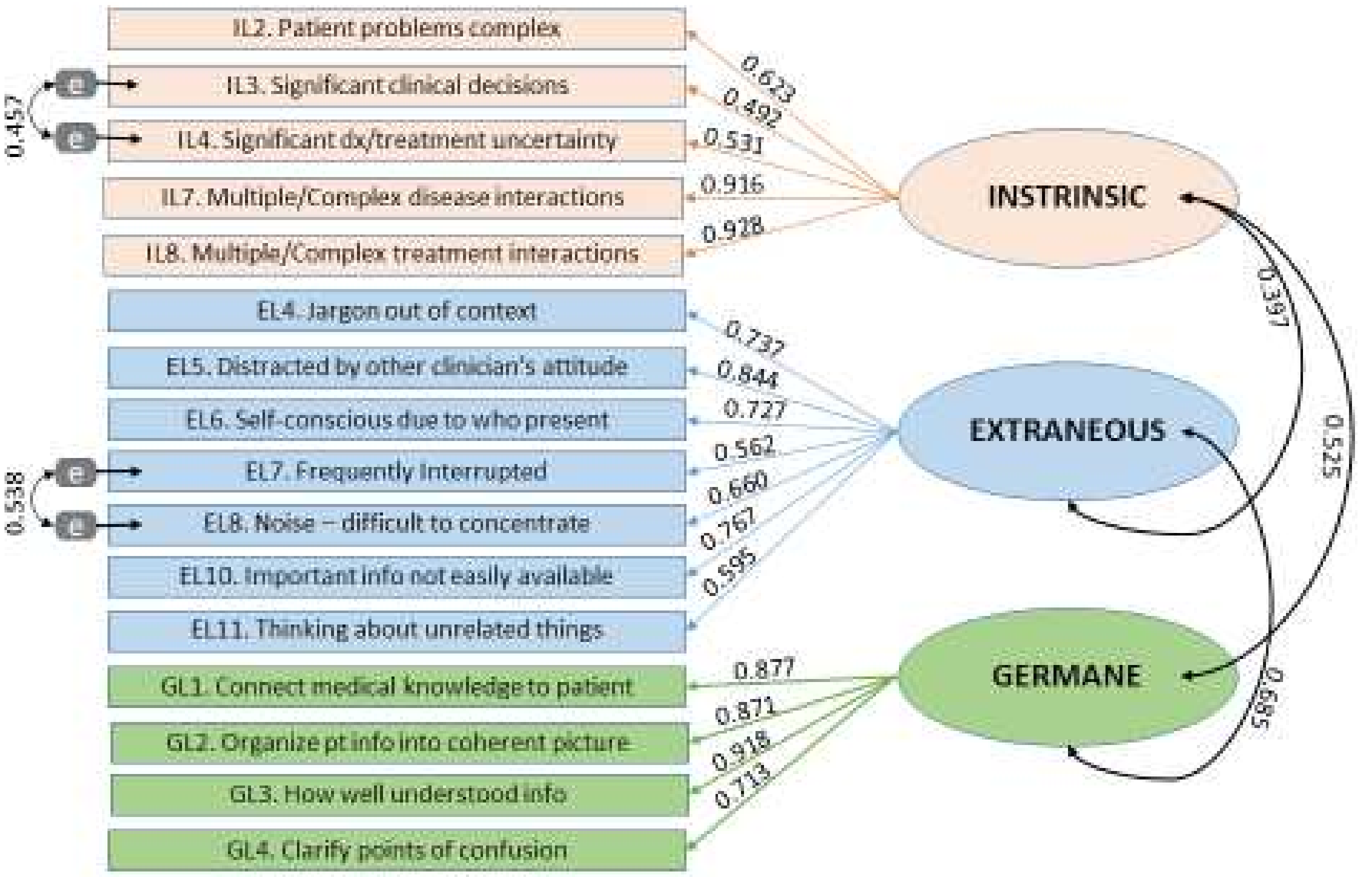
[†] CFI is an estimate of the proportion of sample information explained by the model, and can range from 0 to 1; values above 0.90 are generally considered adequate.

[ϕ] A Tucker-Lewis Index of .95, indicates the model of interest improves the fit by 95% relative to the null model. This index is preferable for smaller samples. Values above 0.80 are acceptable.

[‡] RMSEA indicates how well the model fits with the population covariance matrix. The recommended cut-off point for the RMSEA has varied over the years. In the past a value of ≤ 0.1 was considered acceptable, whereas some scholars have recently proposed a more stringent cut-off of ≤ 0.06.

[§] SRMR is the standardized difference between observed and predicted correlations; a value < 0.06 is considered a good fit.

Figure 1. Path Diagram for the Cognitive Load Inventory for Handoffs


IL2. Patient problems complex
IL3. Significant clinical decisions
IL4. Significant dx/treatment uncertainty
IL7. Multiple/Complex disease interactions
IL8. Multiple/Complex treatment interactions
EL4. Jargon out of context
EL5. Distracted by other clinician's attitude
EL6. Self-conscious due to who present
EL7. Frequently Interrupted
EL8. Noise – difficult to concentrate
EL10. Important info not easily available
EL11. Thinking about unrelated things
GL1. Connect medical knowledge to patient
GL2. Organize pt info into coherent picture
GL3. How well understood info
GL4. Clarify points of confusion
e
e
e
e
0.457
0.538
0.623
0.492
0.531
0.916
0.928
0.737
0.844
0.727
0.562
0.660
0.767
0.595
0.877
0.871
0.918
0.713
INSTRINSIC
EXTRANEOUS
GERMANE
0.397
0.525
0.685

Single headed arrows are factor loadings; double headed arrows are standardized correlation coefficients.

Table 4. Association of Cognitive Load Types with Level of Training and Clinical Setting[a]

	Intrinsic Load	Extraneous Load	Germane Load

	beta (SD), p-value	**beta (SD), p-value**	**beta (SD), p-value**
Level of Training[b]			
R1 versus all other trainees	-0.12 (0.17), 0.49	-0.38 (0.15), 0.01	-1.21 (0.19), <0.0001
Clinical Setting[c]			
All other settings versus ICU	1.18 (0.24), <0.0001	0.06 (0.21), 0.78	0.33 (0.27), 0.22

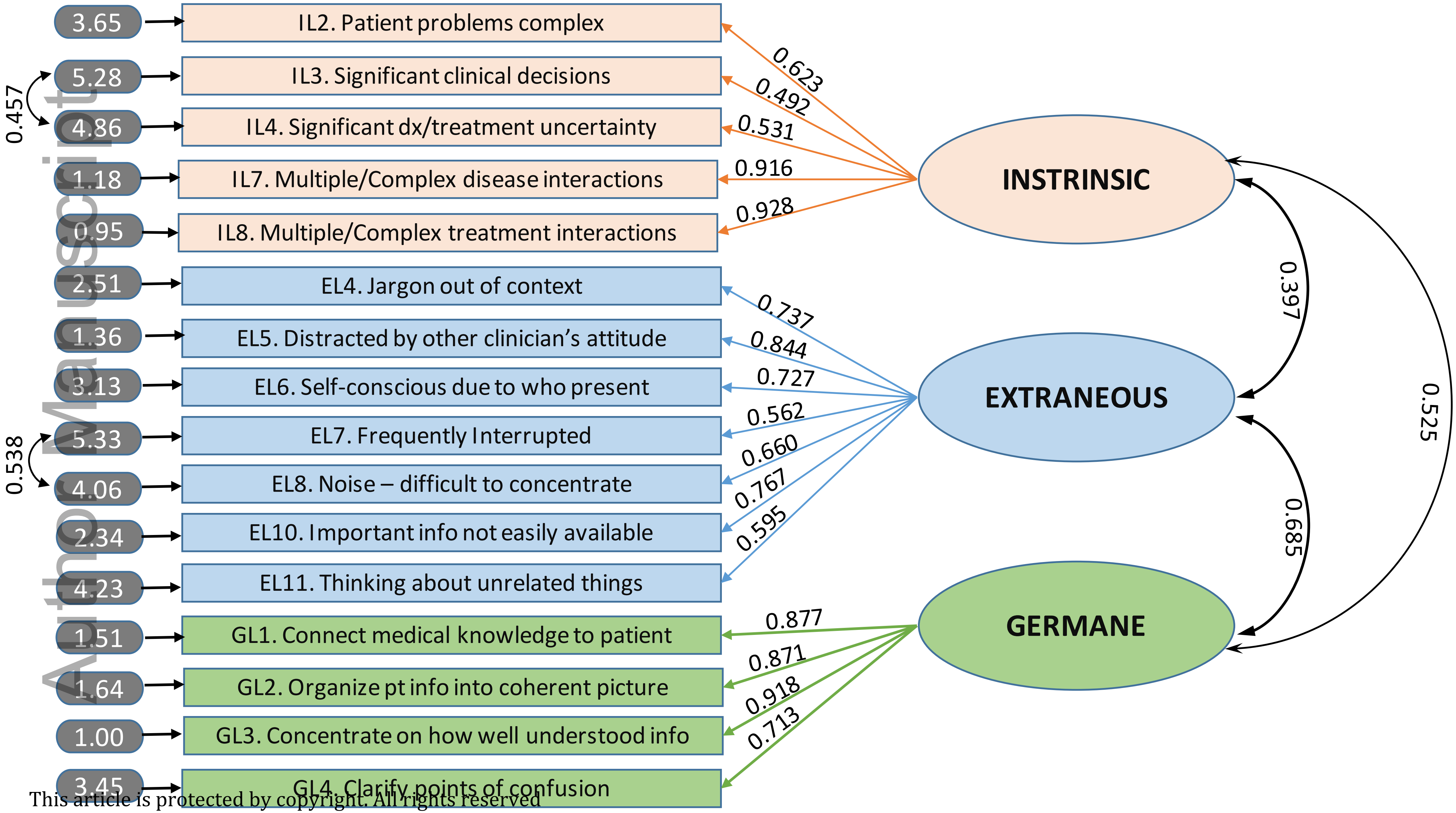
a. beta calculated from univariate regression analysis

b. Our apriori hypothesis was that IL, EL and GL would all decrease as level of training increases.

c. Our apriori hypothesis was that only IL would increase in the ICU compared to other setting.

medu_14292_f1.pptx

Figure 1. Path Diagram for the Cognitive Load Inventory for Handoffs


3.65
IL2. Patient problems complex
5.28
IL3. Significant clinical decisions
4.86
IL4. Significant dx/treatment uncertainty
1.18
IL7. Multiple/Complex disease interactions
0.95
IL8. Multiple/Complex treatment interactions
2.51
EL4. Jargon out of context
1.36
EL5. Distracted by other clinician's attitude
3.13
EL6. Self-conscious due to who present
5.33
EL7. Frequently Interrupted
4.06
EL8. Noise – difficult to concentrate
2.34
EL10. Important info not easily available
4.23
EL11. Thinking about unrelated things
1.51
GL1. Connect medical knowledge to patient
1.64
GL2. Organize pt info into coherent picture
1.00
GL3. Concentrate on how well understood info
3.45
GL4. Clarify points of confusion
0.457
0.538
0.623
0.492
0.531
0.916
0.928
0.737
0.844
0.727
0.562
0.660
0.767
0.595
0.877
0.871
0.918
0.713
INSTRINSIC
EXTRANEOUS
GERMANE
0.397
0.685
0.525

Single headed arrows on the left are error terms and on the right are factor loadings; double headed arrows are standardized correlation coefficients.