



COGNITIVE LOAD

Evidence for validity for the Cognitive Load Inventory for Handoffs

John Q. Young^{1,2}  | Majnu John³ | Krima Thakker⁴ | Karen Friedman⁵  |
Rebekah Sugarman⁶ | Justin L. Sewell⁷  | Patricia S. O'Sullivan⁷

¹Department of Psychiatry, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, New York, USA

²Department of Psychiatry, Zucker Hillside Hospital at Northwell Health, Glen Oaks, New York, USA

³Division of Research, Zucker Hillside Hospital at Northwell Health, Glen Oaks, New York, USA

⁴Division of Education and Training, Zucker Hillside Hospital at Northwell Health, Glen Oaks, New York, USA

⁵Department of Medicine, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, New York, USA

⁶University of Michigan School of Medicine, Ann Arbor, Michigan, USA

⁷Department of Medicine, University of California, San Francisco School of Medicine, San Francisco, California, USA

Correspondence

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

Abstract

Context: Patient handovers remain a significant patient safety challenge. Cognitive load theory (CLT) can be used to identify the cognitive mechanisms for handover errors. The ability to measure cognitive load types during handovers could drive the development of more effective curricula and protocols. No such measure currently exists.

Methods: The authors developed the Cognitive Load Inventory for Handoffs (CLIH) using a multi-step process, including expert interviews to enhance content validity and talk-alouds to optimise response process validity. The final version contained 28 items. From January to March 2019, we administered a cross-sectional survey to 1807 residents and fellows from a large health care system in the USA. Participants completed the CLIH following a handover. Exploratory factor analysis of data from one-third of respondents identified high-performing items; confirmatory factor analysis of data from the remaining sample assessed model fit. Model fit was evaluated using the comparative fit index (CFI) (>0.90), Tucker-Lewis index (TFI) (>0.80), standardised root mean square residual (SRMR) (<0.08) and root mean square of error of approximation (RMSEA) (<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 (five for intrinsic load, seven for extraneous load and four for germane load) supported a three-factor model and met criteria for good model fit: the CFI was 0.95, TFI was 0.93, RMSEA was 0.074 and SRMR was 0.07. The factor structure was comparable for gender and role. Intrinsic, extraneous and germane load scales had high internal consistency. With one exception, scale scores were associated, as hypothesised, with postgraduate level and clinical setting.

Conclusions: The CLIH measures three types of cognitive load during patient handovers. Evidence for validity is provided for the CLIH's content, response process, internal structure and association with other variables. This instrument can be used to determine the relative drivers of cognitive load during handovers in order to optimize handover instruction and protocols.

1 | INTRODUCTION

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

Despite these advances, handovers remain a significant patient safety challenge, even in those studies reporting improvements.¹¹ Researchers have identified cognitive load theory (CLT) as a framework through which to explore the cognitive mechanisms for handover errors and thereby develop more effective handover protocols and curricula.^{12,13} Originally developed by John Sweller in the context of studying how students problem solve,¹⁴ CLT focuses on the implications of limited working memory (WM) for learning.¹⁵ Whereas sensory and long-term memory are relatively infinite, WM is finite. In fact, WM can only actively process (ie, organise, compare and contrast) two to four elements at any given moment.^{16,17} When the cognitive load of a learning task such as a handover exceeds the WM capacity of the trainee, errors occur, often in the form of information loss (eg, about drug allergy, critical co-morbidity, relevant history or current treatments) or distortion (eg, wrong medication dose, wrong surgical site, incorrect diagnosis). This can lead to patient harm.

Cognitive load theory envisions at least two, and possibly three, types of cognitive load.¹⁸ *Intrinsic load* (IL) arises from the information-processing demands associated with the performance of the task itself. Both task complexity and learner expertise determine the IL imposed by any given handover. *Extraneous load* (EL) occurs when learners use WM resources to process information not essential to the task. Common causes include external distractions (eg, noise in the environment) or suboptimal instructional design (eg, unnecessary need to search for information).¹⁹ Internal distractions (eg, worries about external or personal issues, competing demands, self-induced time pressure) may also contribute to EL.^{20,21} *Germane load* (GL) 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 GL may be best understood as a component of IL rather than as a separate type of load.²³⁻²⁵ Yet, some empirical work has found evidence for GL as a separate type.^{26,27}

In order to identify the cognitive mechanisms of current handover protocols and to develop new handover strategies that modulate IL, EL and GL in the desired directions, we need measures that

Key message

This article establishes evidence for validity of the Cognitive Load Inventory for Handoffs, a self-report instrument that measures cognitive load types during patient handovers. This instrument can help drive improvements in handover instruction and protocols.

differentiate between the types of cognitive load that impact a learner during a handover. To date, two studies have attempted to develop a measure of cognitive load types during handovers.^{27,28} The EL items in both studies performed poorly and the findings related to GL were contradictory. To build upon this prior work, we revised the Cognitive Load Inventory for Handoffs (CLIH) and then collected evidence for its validity.

2 | METHODS

2.1 | Study design and ethics

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

The Institutional Review Board for Northwell Health reviewed the study protocol and deemed the study to be exempt from requirements for ethical approval as an educational research study with minimal risk.

2.2 | Development of the CLIH

We followed recommendations in the literature for survey development.³² We reviewed previously published literature on CLT and medical education,^{18,33} drivers of cognitive load during handovers,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 handover feature (Appendix S1). For example, an IL item asked about the difficulty of managing the amount of clinical information. This item was mapped to the CLT construct of number of information elements and to handover features such as the number of patients, co-morbidities 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 handover features of uncertainty (eg, diagnostic), interactions (eg, drug-drug) and maturity of the evidence base for the disease. The first author then conducted individual interviews with nine international experts in, respectively, CLT ($n = 5$) and handovers ($n = 4$). These experts provided feedback on each item with respect to its clarity and alignment with the intended CLT construct and handover feature. Items were revised, deleted or added after each interview. The adaptation of items from previously published work, development of additional items to ensure representativeness to the domains of CLT and handovers, and expert review enhanced item quality and content evidence of validity.

We conducted cognitive interviews with trainees in order to reduce construct-irrelevant variance and optimise the response process (ie, the probability that the resident interpreted each item as intended by its authors). First, we (JQY and RS) performed interviews with two groups (five chief residents in psychiatry and 12 third-year residents in psychiatry). Then we conducted individual interviews with seven internal medicine and two paediatric residents. All of these residents were identified by their respective programme directors and agreed to participate. During the group and individual interviews, each item was read and residents were asked to articulate aloud what they thought the item was asking. When JQY 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 11-point scale (0 = strongly disagree, 10 = strongly agree). Ten items measured IL, 12 items measured EL and six items measured GL. We intentionally included a larger number of items than we had intended to ultimately retain in order to minimise construct under-representation, especially for EL, which two prior studies had failed to measure.^{27,28} We also developed three global rating items, one for each cognitive load type, for internal validation. Appendix S2 shows the 28 CLIH items and three global items.

2.3 | Participants and procedures

We recruited residents and fellows from Northwell Health, a large, 24-hospital health care system in the New York City metropolitan area, which sponsors 122 distinct Accreditation Council for Graduate Medical Education-accredited residency and fellowship programmes. The Office of Academic Affairs provided email addresses for the 1823 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 of training, specialty of training, service and setting in which the handover had occurred, reason for the handover, and number of hours since the handover had been completed. Non-respondents were sent weekly emails over 7 weeks in order to increase the response rate.³⁶ Reminders were sent at different times of day (06.00 hours, 09.00 hours, 12.00 hours and 17.00 hours) to capture the transitions between night-day and day-night shifts. We asked participants to complete the survey after a handover. Invitees could participate only once and could enter a draw for one of four US\$250 gift cards.³⁷

2.4 | Outcomes and analysis

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

2.4.1 | 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 handover constructs, iterative revisions based on input from relevant experts, and the inclusion of more items than were expected to be retained in order to minimise construct under-representation. The validity of the response process was enhanced by the multiple cognitive interviews, which helped to identify sources of confusion in the items and thereby reduce construct-irrelevant variance.

2.4.2 | 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. Firstly, prior versions of the tool had not performed well, especially the items intended to measure EL and GL, leading to one-factor and two-factor solutions in two different studies.^{27,28} Secondly, although we hypothesised three factors, strong arguments, both theoretical and empirical, have been made for both two-factor and three-factor models.²⁵ Finally, we added emotion items to measure EL, which had not been used before in the measurement of cognitive load. We were not sure if the emotion items would map on to a single construct of EL or lead to two different EL factors. Given these uncertainties about the factor structure of the CLIH, we pursued a two-step process. In step 1, we conducted exploratory factor analysis (EFA) to assess the performance of individual items and to better understand the factor structure. In step 2, we performed confirmatory factor analysis (CFA) to cross-validate the factor structure and evaluate model fit.

We used a split-sample strategy.^{38,39} One-third of the total sample was randomly assigned to EFA; there was no overlap between the samples used for EFA and CFA. Consistent with expert recommendations, data from the 11-point CLIH scale were treated as interval and therefore parametric methods were used.^{40,41} Ordinary least squares analysis was employed for EFA; this approach produces solutions very similar to maximum likelihood even when the underlying matrices are badly behaved. All EFAs were conducted using the function 'fa' within the 'psych' package in R (R Foundation for Statistical Computing, Vienna, Austria). In order to allow the items to more distinctly group into a factor, Varimax rotation, which maximises the sum of the variance of the squared loadings, was applied to the minimum residual solution. Exploratory factor analysis was performed iteratively. To be included, a factor was required to have an eigenvalue of >1 and to contain at least two items with loadings of >0.40. At each iteration an item was removed if the item was split across factors or if the corresponding factor loading was <0.40.

We performed CFA using the PROC CALIS procedure in SAS Version 9.4 (SAS Institute, Inc., Cary, NC, USA). Model fit was evaluated using the comparative fit index (>0.90), Tucker-Lewis index (>0.80), standardised root mean square residual (<0.08), and root mean square of approximation (<0.08).^{42,43} Confirmatory factor analysis generates standardised 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. Firstly, 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. Secondly, to assess internal consistency, we examined Cronbach's alpha. Finally, in order to assess validity and generalisability across various sub-populations, we utilised multi-group CFA. We divided the sample by gender and by role (patient handover information sender versus patient handover information receiver) and conducted separate subgroup CFAs. 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-squared values between the unconstrained model and the weight-constrained model was above the 5% significance level.

2.4.3 | Relationships to other variables

We used two methods to assess relationships to other variables. Firstly, we used Pearson's *r* to examine the correlations between the global rating items and total IL, EL and GL. Secondly, we used univariate regression to analyse how each cognitive load type varied with level of training and with clinical setting. Because we hypothesised the most significant difference to be between interns and all others, we dichotomised the respondents accordingly. Similarly, we dichotomised clinical setting into intensive care unit (ICU) versus other settings because we hypothesised 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.

3 | RESULTS

Of the 1823 trainees invited to participate, 16 had email addresses to which email was undeliverable, which resulted in a pool of 1807 potential participants. We received 693 responses (38.4%), representing all training programmes in the health care system. A total of 231 responses were randomly assigned to EFA and 462 to CFA. Table 1 summarises the characteristics of the respondents. The majority of the respondents were in their first 3 years of residency (78%); males and females were equally represented. Most handovers had taken place at the end of a shift (79%) and had occurred in the non-ICU in-patient setting (67%). Participants came from all specialties, with the majority representing non-surgical disciplines (77%). The average number of patients per handover was 10.3, but a large standard deviation (SD) of 10.5 indicated 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 proportions of respondents on a surgical service (greater proportion in the EFA sample) and in the proportion of respondents in a non-surgical residency programme (greater proportion in the CFA sample). Mean IL, EL, and GL did not differ between the two groups. About 27% of respondents replied to reminders/invitations 4-7 (roughly 4-7 weeks after the initial invitation).

3.1 | Internal structure

3.1.1 | 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 two GL items (GL5, GL6) were removed sequentially because they performed poorly (ie, factor loadings of <0.40 and/or were split across factors). The final model had 16 items (five IL, seven EL and four GL items) and produced a three-factor (eigenvalues of >1) model explaining 52% of the total variance (Appendix S2). Item loadings were high (0.52-0.90 for IL, 0.40-0.75 for EL, 0.50-0.86 for GL) and only one cross-loading was higher than 0.3.

3.1.2 | Confirmatory factor analysis

Fifty-one of the 462 surveys assigned to the CFA were incomplete and were excluded from the analysis. All factor loadings were well above 0.50 and were statistically significant (Table 2). Overall mean \pm SD values were 4.76 ± 2.06 for IL, 2.65 ± 1.88 for EL and 3.45 ± 2.29 for GL. Modification indices identified intercorrelation between two pairs of items (IL3/IL4 and EL7/EL8). We allowed these two pairs to

TABLE 1 Characteristics of participants

| Characteristic | Total (n = 693) | | EFA (n = 231) | | CFA (n = 462) | | P-value for EFA vs CFA ^a |
|---|-----------------|-------|---------------|-------|---------------|-------|-------------------------------------|
| | n | % | n | % | n | % | |
| Year of training, residents | | | | | | | |
| PGY-1 | 215 | 31.02 | 83 | 35.93 | 132 | 28.57 | .06 |
| PGY-2 | 180 | 25.97 | 50 | 21.65 | 130 | 28.14 | .08 |
| PGY-3 | 144 | 20.78 | 51 | 22.08 | 93 | 20.13 | .63 |
| PGY-4 | 50 | 7.22 | 19 | 8.23 | 40 | 8.66 | .95 |
| Year of training, fellows | | | | | | | |
| PGY-4 | 29 | 4.18 | 6 | 2.60 | 14 | 3.03 | .93 |
| PGY-5 and higher | 74 | 10.68 | 22 | 9.52 | 52 | 11.26 | .57 |
| Missing data | 1 | 0.14 | 0 | 0.00 | 1 | 0.22 | 1.0 |
| Gender | | | | | | | |
| Male | 344 | 49.64 | 104 | 45.02 | 240 | 51.95 | .10 |
| Female | 343 | 49.49 | 124 | 53.68 | 219 | 47.40 | .15 |
| Other | 2 | 0.29 | 2 | 0.87 | 0 | 0.00 | .21 |
| Prefer not to answer | 3 | 0.43 | 1 | 0.43 | 2 | 0.43 | 1.0 |
| Missing data | 1 | 0.14 | 0 | 0.00 | 1 | 0.22 | 1.0 |
| Clinical setting in which the handover occurred | | | | | | | |
| In-patient ICU | 90 | 12.99 | 26 | 11.26 | 64 | 13.85 | .41 |
| In-patient non-ICU | 463 | 66.67 | 160 | 69.26 | 303 | 65.37 | .35 |
| Emergency department | 67 | 9.67 | 21 | 9.09 | 46 | 9.96 | .83 |
| Ambulatory | 28 | 4.04 | 9 | 3.90 | 19 | 4.11 | 1.0 |
| Perioperative setting | 24 | 3.46 | 9 | 3.90 | 15 | 3.25 | .82 |
| Other ^b | 13 | 1.88 | 3 | 1.30 | 10 | 2.16 | .62 |
| Missing data | 9 | 1.30 | 3 | 1.30 | 6 | 1.30 | 1.0 |
| Reason for the handover | | | | | | | |
| End of shift | 550 | 79.37 | 182 | 78.79 | 368 | 79.65 | .86 |
| Transfer to a different team within the same setting ^c | 35 | 5.05 | 12 | 5.19 | 23 | 4.98 | 1.0 |
| Transfer to a different setting ^d | 36 | 5.19 | 14 | 6.06 | 22 | 4.76 | .59 |
| End of rotation | 61 | 8.80 | 19 | 8.23 | 42 | 9.09 | .81 |
| Other ^e | 2 | 0.29 | 1 | 0.43 | 1 | 0.22 | 1.0 |
| Missing data | 9 | 1.30 | 3 | 1.30 | 6 | 1.30 | 1.0 |
| Number of hours since completion of the handover | | | | | | | |
| 0-24 h | 366 | 52.81 | 119 | 51.52 | 247 | 53.46 | .60 |
| 24 h to 5 days | 119 | 17.17 | 49 | 21.21 | 70 | 15.15 | .07 |
| >5 days | 204 | 29.44 | 63 | 27.27 | 141 | 30.52 | .39 |
| Missing | 4 | 0.58 | 0 | 0.00 | 4 | 0.87 | .38 |
| Specialty in which the handover occurred | | | | | | | |
| Surgical ^f | 148 | 21.36 | 60 | 25.97 | 88 | 19.05 | .04 |
| Non-surgical ^g | 533 | 76.91 | 168 | 72.73 | 365 | 79.00 | .06 |
| Other | 4 | 0.58 | 1 | 0.43 | 3 | 0.65 | .54 |
| Missing data | 8 | 1.15 | 2 | 0.87 | 6 | 1.30 | .90 |
| Specialty of the trainee | | | | | | | |
| Surgical ^h | 146 | 21.07 | 59 | 25.54 | 87 | 18.83 | .05 |
| Non-surgical ⁱ | 527 | 76.05 | 163 | 70.56 | 364 | 78.79 | .02 |

(Continues)

TABLE 1 (Continued)

| Characteristic | Total (n = 693) | | EFA (n = 231) | | CFA (n = 462) | | P-value for EFA vs CFA ^a |
|--|-----------------|-------|-------------------|-------|-------------------|-------|-------------------------------------|
| | n | % | n | % | n | % | |
| Other (transitional year) | 17 | 2.45 | 8 | 3.46 | 9 | 1.95 | .34 |
| Missing data | 3 | 0.43 | 1 | 0.43 | 2 | 0.43 | 1.0 |
| Role in handover | | | | | | | |
| Sender | 559 | 80.66 | 178 | 77.06 | 381 | 82.47 | .10 |
| Receiver | 125 | 18.04 | 50 | 21.65 | 75 | 16.23 | .10 |
| Missing data | 9 | 1.30 | 3 | 1.30 | 6 | 1.30 | 1.0 |
| Number of patients per handover, mean \pm SD | 38 \pm 10.56 | | 10.66 \pm 11.06 | | 10.25 \pm 10.31 | | 0.64 |

Abbreviations: CFA, confirmatory factor analysis; EFA, exploratory factor analysis; ICU, intensive care unit; PGY, postgraduate year; SD, standard deviation.

^aP-values are for chi-squared tests for proportions and t-tests for means.

^bCall room, lecture, phone, resident quarters, office, between shifts.

^cFor example, transfer from surgery to medicine.

^dFor example, transfer from in-patient to out-patient.

^eAfternoon rounding, communication between team members.

^fAnaesthesiology, general surgery, neurosurgery, obstetrics and gynaecology, oral surgery, orthopaedics, urology, vascular surgery.

^gCardiology, critical care, dental medicine, dermatology, emergency medicine, endocrinology, ear/nose/throat, family medicine, gastroenterology, haematology/oncology, internal medicine, nephrology, neurology, neonatal ICU, ophthalmology, paediatrics, paediatric infectious disease, palliative care, physical medicine and rehabilitation, podiatry, pulmonary medicine, psychiatry, radiology, radiation oncology, rheumatology, surgery ICU.

^hAnaesthesiology, general surgery, neurological surgery, obstetrics and gynaecology, oral surgery, orthopaedic surgery, plastic surgery, thoracic surgery, urology, vascular surgery.

ⁱDermatology, emergency medicine, family medicine, internal medicine, neurology, neuroradiology, ophthalmology, oral pathology, pathology, paediatrics, paediatric dental medicine, physical medicine and rehabilitation, podiatry, psychiatry, radiology, radiation oncology.

correlate. In the modified three-factor model, the goodness-of-fit parameters were all favourable and exceeded our predetermined thresholds (Table 3). The correlations between each factor were moderate: $r_{il,el} = .40$, $r_{il,gl} = .52$ and $r_{el,gl} = .68$ (Table 2). Figure 1 depicts the path diagram for the measurement model and highlights the factor structure, factor loadings and correlations between the factors.

3.1.3 | Additional internal structure evidence

The goodness-of-fit parameters were superior for the three-factor model compared with the two-factor model, suggesting that GL was a separate source of variance and not nested within IL (Table 3). The 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, the factor structure was stable across the subgroups for gender and role; there was no difference between the unconstrained models and weight-constrained model for gender (χ^2 difference = 9.05, $df = 13$, P -value = .77) and role (sender versus receiver; χ^2 difference = 17.93, $df = 13$, P -value = .16).

3.2 | Relationships 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 with interns. However, IL for more advanced trainees did not differ from that for interns. The relationship between cognitive load types and clinical setting confirmed the a priori hypotheses. Intrinsic load was significantly higher in the ICU compared with other settings, whereas there were no differences for EL and GL (Table 4).

4 | DISCUSSION

In this study, we developed and tested an instrument designed to measure the cognitive load experienced by trainees during patient handovers. 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 constructs and handover features and multiple cognitive interviews with trainees. The results marshal strong evidence supporting the internal structure of the CLIH. Exploratory factor analysis produced a three-factor structure with 16 high-performing items. Confirmatory factor analysis confirmed the superiority of a three-factor model compared with a two-factor model with excellent fit indices. In addition, internal consistency was high and the factor structure did not differ between female and male respondents or senders and receivers. Finally, with a few exceptions that will 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 IL was significantly higher in the ICU. Taken together, these findings are very

TABLE 2 Confirmatory factor analysis results

| Item | Item | Mean \pm SD | Factor loading | Standard error | P-value |
|--|---|-----------------|----------------|----------------|---------|
| Intrinsic load: Please rate your agreement with the following statements regarding the handover you have completed: | | | | | |
| IL2 | The patient problems were complex | 5.34 \pm 2.45 | 0.62 | 0.03 | <.0001 |
| IL3 ^a | The handover included significant clinical decision(s) that needed to be made | 4.99 \pm 2.65 | 0.49 | 0.04 | <.0001 |
| IL4 ^a | The handover included significant diagnostic and/or treatment uncertainty | 4.23 \pm 2.60 | 0.53 | 0.04 | <.0001 |
| IL7 | I had to consider multiple or complex interactions between diseases | 4.54 \pm 2.71 | 0.92 | 0.01 | <.0001 |
| IL8 | I had to consider multiple or complex interactions between treatments | 4.50 \pm 2.62 | 0.93 | 0.01 | <.0001 |
| Global IL | Overall, I found the patient problems difficult to understand | - | - | - | - |
| | Overall mean for IL ^{b,c,d} | 4.76 \pm 2.06 | | | |
| Extraneous load: Please rate your agreement with the following statements regarding the handover. These statements are about the environment and your mindset during the handover: | | | | | |
| EL4 | The other clinician used jargon out of context | 2.37 \pm 2.35 | 0.74 | 0.03 | <.0001 |
| EL5 | I was distracted by the other clinician's attitude | 1.96 \pm 2.20 | 0.84 | 0.02 | <.0001 |
| EL6 | I was self-conscious due to who was present | 2.41 \pm 2.60 | 0.73 | 0.03 | <.0001 |
| EL7 ^a | I was frequently interrupted (eg, pages, phone calls, people, etc.) | 3.37 \pm 2.79 | 0.56 | 0.04 | <.0001 |
| EL8 ^a | Noise made it difficult to concentrate | 3.00 \pm 2.68 | 0.66 | 0.03 | <.0001 |
| EL10 | During the handover, important information was not easily available when I needed it | 2.56 \pm 2.39 | 0.77 | 0.02 | <.0001 |
| EL11 | I was thinking about things unrelated to the sign-out | 2.89 \pm 2.55 | 0.60 | 0.04 | <.0001 |
| Global EL | Overall, I found it difficult to focus my attention on the handover | - | - | - | - |
| | Overall mean for EL ^{b,c,d} | 2.65 \pm 1.88 | | | |
| Germane load: Please rate your agreement with the following statements regarding your mental effort during the handover you have completed: | | | | | |
| GL1 | I had to work hard to connect my own medical knowledge to the patient problems | 3.18 \pm 2.56 | 0.88 | 0.01 | <.0001 |
| GL2 | I had to work hard to organise the patient information into a coherent clinical picture | 3.49 \pm 2.61 | 0.87 | 0.01 | <.0001 |
| GL3 | During the sign-out, I had to work hard to concentrate on how well I understood the information | 3.11 \pm 2.53 | 0.92 | 0.01 | <.0001 |
| GL4 | I had to take steps to clarify points of confusion | 3.95 \pm 2.65 | 0.71 | 0.03 | <.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 \pm 2.29 | | | |

Abbreviations: EL, extraneous load; GL, germane load; IL, intrinsic load; SD, standard deviation.

^aTwo pairs of items (IL3/IL4 and EL7/EL8) were allowed to correlate.

^bOverall mean = sum of the items answered divided by the number of items answered; 11-point scale (0-10, strongly disagree to strongly agree).

^cCronbach's alpha: IL = 0.85; EL = 0.87, and GL = 0.91.

^dCorrelations between scales: $r_{il,el} = .40$; $r_{il,gl} = .52$, and $r_{el,gl} = .68$.

encouraging and support the ability of the CLIH to measure IL, EL and GL during handovers. With such a measure, educators can study tailored handover interventions.

Although the overall findings provide strong support for the CLIH, there were two specific results that were not expected. Firstly, IL did not decrease for more advanced learners. When we examined mean IL for each level of training, it appeared that IL decreases modestly from intern year until the end of residency but then increases during the first year of fellowship. This finding is likely to reflect how the transition from senior resident to first-year fellow puts the

trainee in a new role and setting in which 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 increasing expertise is matched with more challenging roles that keep IL more or less constant. The second inconsistent finding relates to the low correlation between the global rating item for GL and the GL score. In retrospect, we think this may relate to poor construction of the GL global rating item. The focus on 'activities to better understand

TABLE 3 Measures of fit for two-factor and three-factor models

| Model | χ^2 , <i>df</i> , <i>P</i> -value, Normed χ^2 ^a | CFI ^b | Tucker-Lewis index ^c | RMSEA ^d (95% CI) | SRMR ^e |
|-------------------------------|---|------------------|---------------------------------|-----------------------------|-------------------|
| Two-factor model | $\chi^2 = 1154.7$, <i>df</i> = 103, <i>P</i> < .0001, Normed $\chi^2 = 11.2$ | 0.74 | 0.70 | 0.158 (0.150, 0.166) | 0.1026 |
| Two-factor model (modified) | $\chi^2 = 907.3$, <i>df</i> = 101, <i>P</i> < .0001, Normed $\chi^2 = 5.32$ | 0.80 | 0.76 | 0.139 (0.1313, 0.1479) | 0.0993 |
| Three-factor model | $\chi^2 = 537.9$, <i>df</i> = 101, <i>P</i> < .0001, Normed $\chi^2 = 5.32$ | 0.89 | 0.87 | 0.103 (0.0943, 0.1113) | 0.0754 |
| Three-factor model (modified) | $\chi^2 = 322.363$, <i>df</i> = 99, <i>P</i> < .0001, Normed $\chi^2 = 3.26$ | 0.95 | 0.93 | 0.074 (0.065, 0.083) | 0.0735 |

Abbreviations: CFI, comparative fit index; CI, confidence interval; *df*, degrees of freedom; RMSEA, root mean square error of approximation; SRMR, standardised root mean square residual.

^aA non-significant (*P* > .05) χ^2 -value suggests the model is an adequate representation of the data. However, with a large sample size (>200), the χ^2 -value 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) χ^2 -value is recommended. This value equals the χ^2 index divided by the degrees of freedom. The criterion for acceptance is recommended as <5.

^bCFI 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.

^cA Tucker-Lewis index of 0.95 indicates the model of interest improves the fit by 95% relative to the null model. This index is preferable for smaller samples. Values of >0.80 are acceptable.

^dRMSEA indicates how well the model fits with the population covariance matrix. A value of <0.08 is considered a good fit.

^eSRMR is the standardised difference between observed and predicted correlations; a value <0.08 is considered a good fit.

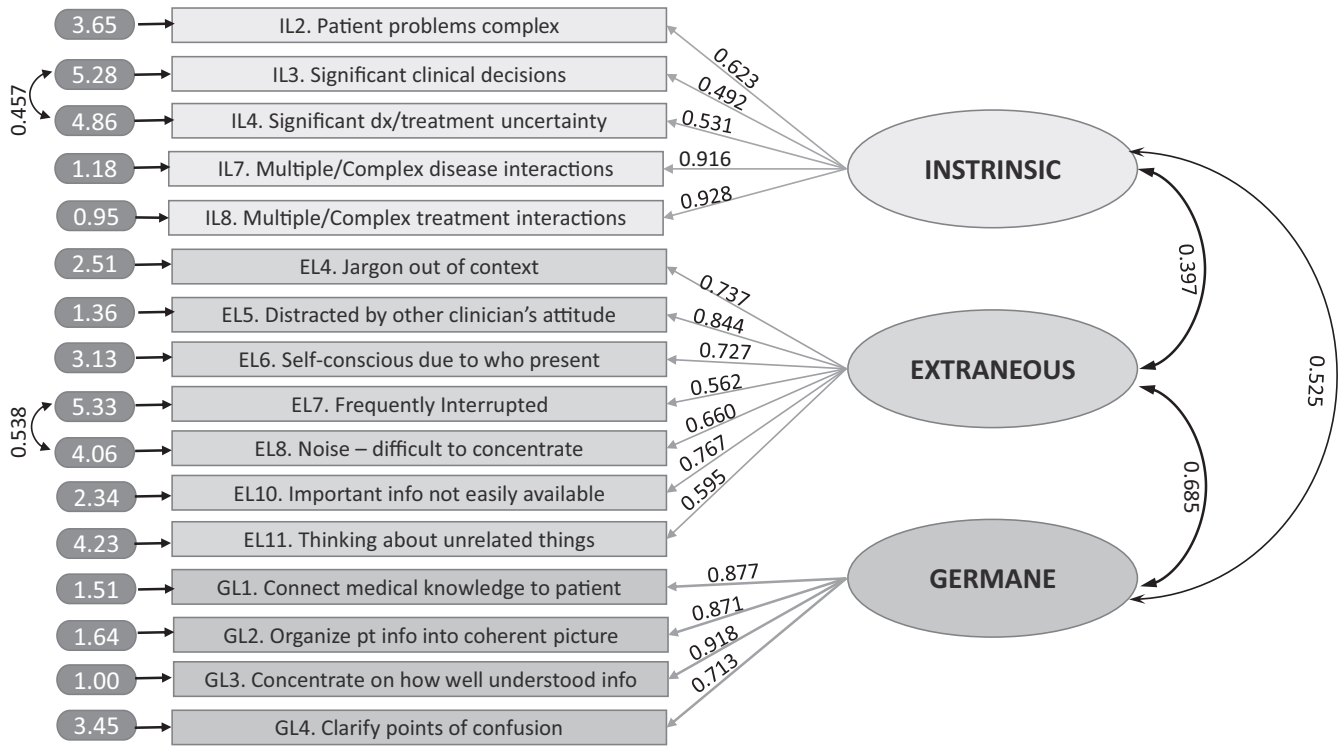
the patient problems' may not adequately differentiate between IL and GL.

This study has several methodological strengths. Although CLT is considered highly applicable to patient handovers, efforts to measure cognitive load have encountered difficulty. Two prior studies successfully measured IL, but both failed to form a stable factor structure for EL and reported inconsistent findings with regard to whether GL was best understood as a separate factor or as nested within IL.^{27,28} The present study's success in measuring the three types of cognitive load may be attributable to three key differences between it and the prior studies. Firstly, item development included systematic mapping to CLT constructs and handover features and iterative revisions based on input from many experts. Secondly, in this study, we engaged in extensive evaluation of the response process through numerous think-aloud exercises with both small groups and individuals. This process enabled us to identify and address multiple instances in which the wording of the item was either confusing to trainees or interpreted in a manner that differed from that intended. Finally, unlike the prior two studies,^{27,28} which measured cognitive load in the context of simulated handovers, the present study asked trainees to complete the instrument after an actual handover. This may have especially influenced the performance of the EL items because both prior studies on handovers occurred in the context of simulations that intentionally removed distractions and other sources of EL.^{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 any of a four-factor model (in which the internal distraction items form a separate factor in addition to IL, EL and GL), a three-factor model (including IL, EL and GL) and 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. Firstly, the internal distraction and interpersonal friction items loaded on to 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 relating mostly to task design and more recently the environment (eg, interruptions).^{19,25} These results suggest that internal distraction (eg, worry or self-consciousness) and interpersonal friction (eg, annoyance with the style of another person) contribute to EL. Future research should examine the extent to which these factors influence learning and performance.

Secondly, researchers currently debate whether GL should be conceptualised as a third type of cognitive load distinct from IL and EL or as a subset of IL.^{20,45-47} Many CLT researchers are now advocating for a two-factor model that understands GL as a component of IL.²⁵ By contrast, this study's results were most consistent with a three-factor model. Interestingly, the only other instrument



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

FIGURE 1 Path diagram for the cognitive load inventory for handovers. Single-headed arrows are factor loadings; double-headed arrows are standardised correlation coefficients [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Association of cognitive load types with level of training and clinical setting^a

| | Intrinsic load beta (SD), P-value | Extraneous load beta (SD), P-value | Germane load beta (SD), P-value |
|--------------------------------|---|---|--|
| Level of training ^b | | | |
| R1 vs all other trainees | -0.12 (0.17), .49 | -0.38 (0.15), .01 | -1.21 (0.19), <.0001 |
| Clinical setting ^c | | | |
| All other settings vs ICU | 1.18 (0.24), <.0001 | 0.06 (0.21), .78 | 0.33 (0.27), .22 |

Abbreviations: ICU, intensive care unit; R1, year-1 resident; SD, standard deviation.

^aBeta calculated from univariate regression analysis.

^bOur a priori hypothesis was that intrinsic, extraneous and germane load would all decrease as level of training increased.

^cOur a priori hypothesis was that only intrinsic load would increase in the ICU compared with other settings.

developed specifically for a medical education context is one that measures cognitive load during colonoscopy.²⁶ That measure also demonstrated three factors. Although these results challenge the two-factor proponents, it is important to note that the third factor in both this and the colonoscopy study²⁶ may represent not GL, but

another construct. Germane load, by definition, enhances learning. Examples of means to promote GL include instructional design (eg, interleaved practice compared with blocked practice) and the prompting of generative processes (eg, self-explanation or elaboration).⁴⁸ Future studies might address this question by evaluating whether learning improves with instructional techniques that impact GL but not IL. Such studies must be careful to differentiate learning from performance by, for example, measuring the impact of a technique several weeks later in a different context.

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

The study has several limitations. It achieved a participant response rate of <40%. We do not know whether non-responders differed from responders. In addition, the study was conducted in a single health care system. However, this single health care system is diverse and participants in the study came from multiple specialties and hospitals. The participants included only residents and fellows and hence we do not know how this instrument might function with students or faculty members. Future studies should evaluate the stability of the factor structure in other populations and settings. Further, the CLIH is based on learner recall after the fact. More than a third of participants completed the survey more than 24 hours after the handover. This is a significant length of time and introduces recall bias. For example, if transient but significant events such as distractions were poorly remembered, 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 subgroup CFA for time between handover and completion of the CLIH (up to 24 hours versus more than 24 hours). There was no difference in the factor structure between the two groups (χ^2 difference = 9.058, $df = 13$, P -value = .769).

In conclusion, the CLIH shows evidence of ability to measure cognitive load types (IL, EL and GL) during patient handovers within a large sample of trainees from multiple specialties and hospitals. Improving handover instruction requires strategies that reduce EL and optimise 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.

AUTHOR CONTRIBUTIONS

JQY, JLS and PSO'S were the primary conceivers of the study. All authors made substantial contributions to the acquisition, analysis and interpretation of the data, and to the drafting and revision of the manuscript. All authors approved the final manuscript for publication and have agreed to be accountable for all aspects of the work.

ACKNOWLEDGEMENTS

None.

CONFLICTS OF INTEREST

None.

ETHICAL APPROVAL

Ethical approval was obtained from the Institutional Review Board of Northwell Health (IRB no. 18-0192).

ORCID

John Q. Young  <https://orcid.org/0000-0003-2219-5657>

Karen Friedman  <https://orcid.org/0000-0003-1980-1839>

Justin L. Sewell  <https://orcid.org/0000-0003-4049-2874>

REFERENCES

- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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 Qual Saf.* 2016;25(2):66-70.
- Sweller J. Cognitive load during problem solving: effects on learning. *Cogn Sci.* 1988;12(2):257-285.
- Sweller J, van Merriënboer JGG. Cognitive load theory and instructional design for medical education. In: Walsh K, ed. *The Oxford Textbook of Medical Education.* Oxford: Oxford University Press; 2013:74-85.
- Baddeley A. Working memory: theories, models, and controversies. *Annu Rev Psychol.* 2012;63(1):1-29.
- 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.
- Young JQ, van Merriënboer J, Durning S, ten Cate O. Cognitive load theory: implications for medical education: AMEE Guide No. 86. *Med Teach.* 2014;36(5):371-384.
- Choi H-H, van Merriënboer JGG, 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.
- Young JQ, Sewell JL. Applying cognitive load theory to medical education: construct and measurement challenges. *Perspect Med Educ.* 2015;4(3):107-109.
- Feldon DF. Cognitive load and classroom teaching: the double-edged sword of automaticity. *Educ Psychol.* 2007;42(3):123-137.
- Sweller J, van Merriënboer JGG, Paas FGWC. Cognitive architecture and instructional design. *Educ Psychol Rev.* 1998;10(3):251-296.
- Leppink J, Paas F, van Gog T, van der Vleuten CPM, van Merriënboer JGG. Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learn Instr.* 2014;30:32-42.

24. Sweller J, Ayres PL, Kalyuga S. *Cognitive Load Theory*. New York, NY: 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. *Perspect Med Educ*. 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 Med*. 2016;4:2050312116682254.
29. Kane MT. Current concerns in validity theory. *J Educ Meas*. 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 Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: AERA Publications; 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 Merriënboer 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. *J Biomed Informat*. 2009;42(2):377-381.
36. Dillman DA, Smyth JD, Christian LM. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons; 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 method, a routine for evaluating a factor structure. *Psychology*. 2018;9(8):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, MA: 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. *Adv Health Sci Educ Theory Pract*. 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. *Electr J Bus Res Methods*. 2008;6(1):53-60.
43. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equat Model*. 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. *Perspect Med Educ*. 2015;4(3):119-127.
46. van Merriënboer JJ, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Med Educ*. 2010;44(1):85-93.
47. Kalyuga S. Cognitive load theory: how many types of load does it really need? *Educ Psychol Rev*. 2011;23(1):1-19.
48. Fiorella L, Mayer RE. Eight ways to promote generative learning. *Educ Psychol Rev*. 2016;28(4):717-741.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Young JQ, John M, Thakker K, et al. Evidence for validity for the Cognitive Load Inventory for Handoffs. *Med Educ*. 2021;55:222-232. <https://doi.org/10.1111/medu.14292>