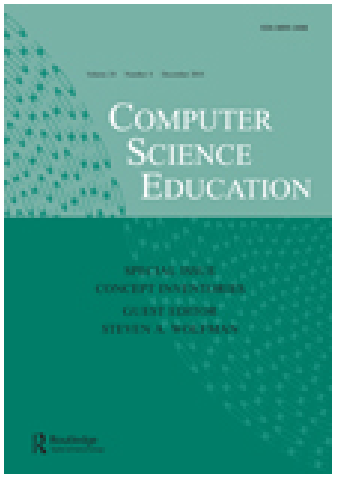


Computer Science Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ncse20

A psychometric evaluation of the digital logic concept inventory

Geoffrey L. Herman[a], Craig Zilles[b] & Michael C. Loui[c]

[a] Illinois Foundry for Innovation in Engineering Education, University of Illinois at Urbana-Champaign, Illinois, IL, USA.

[b] Department of Computer Science, University of Illinois at Urbana-Champaign, Illinois, IL, USA.

[c] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Illinois, IL, USA.

Published online: 24 Oct 2014.

To cite this article: Geoffrey L. Herman, Craig Zilles & Michael C. Loui (2014) A psychometric evaluation of the digital logic concept inventory, Computer Science Education, 24:4, 277-303, DOI: 10.1080/08993408.2014.970781

To link to this article: http://dx.doi.org/10.1080/08993408.2014.970781

Computer Science Education, 2014
Vol. 24, No. 4, 277–303, http://dx.doi.org/10.1080/08993408.2014.970781

A psychometric evaluation of the digital logic concept inventory

Geoffrey L. Herman[a]*, Craig Zilles[b] and Michael C. Loui[c]

[a]*Illinois Foundry for Innovation in Engineering Education, University of Illinois at Urbana-Champaign, Illinois, IL, USA;* [b]*Department of Computer Science, University of Illinois at Urbana-Champaign, Illinois, IL, USA;* [c]*Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Illinois, IL, USA*

(*Received 10 May 2014; accepted 13 September 2014*)

Concept inventories hold tremendous promise for promoting the rigorous evaluation of teaching methods that might remedy common student misconceptions and promote deep learning. The measurements from concept inventories can be trusted only if the concept inventories are evaluated both by expert feedback and statistical scrutiny (psychometric evaluation). Classical Test Theory and Item Response Theory provide two psychometric frameworks for evaluating the quality of assessment tools. We discuss how these theories can be applied to assessment tools generally and then apply them to the Digital Logic Concept Inventory (DLCI). We demonstrate that the DLCI is sufficiently reliable for research purposes when used in its entirety and as a post-course assessment of students' conceptual understanding of digital logic. The DLCI can also discriminate between students across a wide range of ability levels, providing the most information about weaker students' ability levels.

Keywords: concept inventory; digital logic; misconceptions; item response theory; reliability; validity

1. Introduction

A concept inventory is a multiple-choice assessment instrument that measures how well students' conceptual frameworks align with the accepted conceptual frameworks of the target discipline (Hestenes, Wells, & Swackhamer, 1992). Concept inventories have been developed for many science, technology, engineering, and mathematics (STEM) disciplines, consistently revealing that students succeed in traditional classroom assessments through shallow memorization of facts and procedures rather than through the development of deep, conceptual knowledge (Chi, 2006; Evans et al., 2003; Hake, 1998; Hestenes et al., 1992; Litzinger et al., 2010). When students have accurate, deep conceptual knowledge, they can learn more

*Corresponding author. Email: glherman@illinois.edu

efficiently in the future, and they can transfer their knowledge across contexts (Litzinger et al., 2010). Thus, concept inventories have brought attention to a pressing need to develop and adopt pedagogies that better support deep conceptual learning. For example, the Force Concept Inventory (FCI), the first concept inventory, has supported the effectiveness and adoption of active learning pedagogies in physics education (Evans et al., 2003; Hake, 1998, 2002; Mestre, Dufresne, Gerace, Hardiman, & Touger, 1993).

Concept inventories hold tremendous promise if they can indeed measure students' conceptual knowledge. Unfortunately, few concept inventories or similar assessment tools have been scrutinized using measurement or test development theories to justify their use as pedagogy evaluation tools or research instruments (Pellegrino, DiBello, James, Jorion, & Schroeder, 2011; Wallace & Bailey, 2010). In this article, we present a psychometric analysis of the DLCI to ascertain its viability as a research instrument and to make recommendations on its appropriate uses. The article begins by situating our development of the DLCI within the principles of the "assessment triangle." The paper then provides a primer on testing theories and psychometric analysis before applying those theories to the evaluation of the DLCI. Finally, the article concludes with recommendations for the refinement of the DLCI and its appropriate uses.

2. Overview of the development of the DLCI

In the National Research Council publication *Knowing What Students Know: The Science and Design of Educational Assessment*, a panel of assessment experts created the "assessment triangle" as a guiding framework for the rigorous development and evaluation of assessment instruments (Pellegrino, Chudowsky, & Glaser, 2001; Pellegrino, DiBello, & Brophy, 2014). The triangle, as shown in Figure 1, indicates that high quality assessment tools must coordinate and demonstrate three interrelated elements: cognition, observation, and interpretation (Pellegrino et al., 2001).

- *Cognition* refers to a "theory or set of beliefs about how students represent knowledge and develop competence in a subject domain" (Pellegrino et al., 2001, p. 44). As such, an instrument must identify what comprises knowledge in the target domain and have a theory about how students gain expertise in the domain.
- *Observation* "represents a description or set of specifications for assessment tasks that will elicit illuminating responses for students' about the target domain to be measured" (Pellegrino et al., 2001, p. 48). In other words, observations are the tasks that will indicate how far a student has progressed in their trajectory toward expertise as defined by the cognition corner. In the case of a concept inventory,

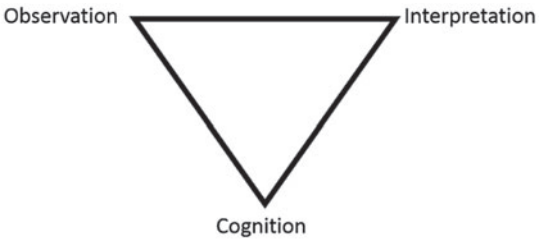


Figure 1. The assessment triangle. Adapted from Pellegrino et al. (2001).

observations are likely to be the presence or absence of common misconceptions.

- *Interpretation* "encompasses all the methods and tools used to reason from fallible observations" (Pellegrino et al., 2001, p. 48). Interpretation acknowledges that students may succeed or fail to perform a task for a variety of reasons that are independent of the ability being measured. For example, differences in notational style (e.g. sign conventions in voltage and current vary between physics and engineering) may lead to unreliable measurements of students' conceptual understanding of a topic. Consequently, interpretation provides tools for understanding the validity and reliability of our measures.

Critically, the development of rigorous assessment tools is an iterative process as engagement with one corner of the assessment triangle informs and refines our understanding of the other corners (Pellegrino et al., 2001; Streveler et al., 2011). Accordingly, the DLCI was developed using two concurrent iterative design cycles grounded in the assessment triangle. The DLCI was refined through expert review feedback cycles (the inner loop in Figure 2) and student performance (the outer loop in Figure 2). As shown in Figure 2, the creation of the DLCI began in the cognition corner, progressed to the observation corner and then the interpretation corner, before iterating through these phases.

2.1. The cognition basis of the DLCI

Because deep, conceptual understanding is vital for future learning and performance, concept inventories use measurements of students' conceptual understanding as a proxy for their mastery of the course content and learning in general (Jorion et al., 2014a). The cognition corner of the assessment triangle for concept inventories asks two basic questions (Streveler et al., 2011): (1) What misconceptions do students have in the domain? and (2) Why do those misconceptions exist and persist? The answer to the second question in particular provides the theoretical underpinning of how the

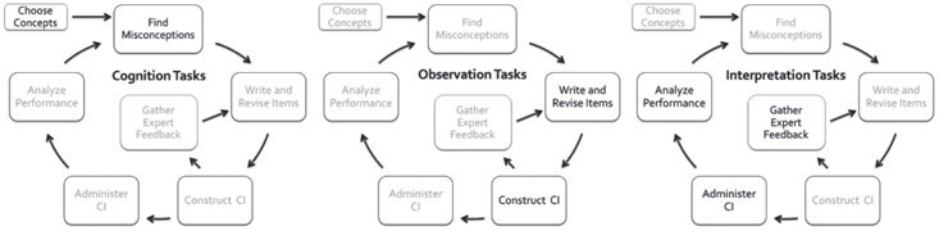


Figure 2. The development cycle uses two concurrent feedback cycles: an inner, expert feedback loop (Write and Revise Items, Construct CI, Gather Expert Feedback) and an outer, student performance feedback loop (Find Misconceptions, Write and Revise Items, Construct CI, Administer CI, Analyze Performance). The DLCI development cycle is shown three times in the figure. Each version highlights which tasks align with the cognition, observation, and interpretation corners of the assessment triangle.

results of the instrument should be observed and interpreted. As we will discuss later, we chose to use DiSessa's theory of Knowledge-in-Pieces (KiP) as the primary basis of the cognition corner of the assessment triangle.

The identification of students' misconceptions in digital logic progressed through a two-staged process of using a panel of experts to identify which topics were important and difficult for students to learn in digital logic (Goldman et al., 2010) and using clinical interviews with students to identify specific misconceptions and their origins (Herman, Loui, Kaczmarczyk, & Zilles, 2012; Herman, Loui, & Zilles, 2011a, 2011b, 2012; Montfort, Herman, Brown, Matusovich, & Streveler, in press; Montfort, Herman, Streveler, & Brown, 2012).

Whereas, there is a consensus on the core content of mechanics, there is no consensus for digital logic (Herman & Loui, 2012). Identifying the core content to be assessed by the DLCI; first required developing a consensus about core content from the community of digital logic instructors and practitioners.

A Delphi process is an iterative survey technique used to moderate dialogue and find consensus among a group of carefully selected experts (Clayton, 1997; Pill, 1971). Clayton (1997), recommends a panel of 15 – 30 experts from a diversity of institutions and backgrounds. For the DLCI, 20 experts were identified based on their authorship of textbooks and rigorous pedagogical articles for digital logic instruction (Goldman et al., 2010). The panel proposed 44 topics, concepts, and skills as essential for students to know after a first course in digital logic and iteratively rated these ideas based on their importance and difficulty. Because the purpose of a concept inventory is to measure a students' conceptual understanding of core concepts, rather than provide a comprehensive assessment, investigations on, and measurements of, students' misconceptions focused on those concepts and skills that were rated as having an importance of eight or higher on a 10-point scale (see Table 1).

Table 1. 15 most important digital logic topics/concepts as identified by the Delphi process.

Concept/skill	Importance
1. *State transitions*: Understanding the difference between the current state and how the current state transitions to the next state	9.8
2. *Converting verbal specifications to state diagrams/tables*	9.8
3. *Functionality of multiplexers, decoders, and other MSI circuits*: Excludes building larger MSI circuits from smaller MSI circuits	9.6
4. *Converting verbal specifications to Boolean expressions*	9.5
5. *Hierarchical design*	9.5
6. *Relating timing diagrams to state machines and circuits*	9.4
7. *Understanding how a sequential circuit corresponds to a state diagram*: Recognizing the equivalence of a sequential circuit and a state diagram	8.9
8. *Modular design*: Building circuits as a compilation of smaller components	8.9
9. *Number representations*: Understanding the relationship between representation and meaning	8.6
10. *Analyzing sequential circuit behavior*	8.5
11. *Converting algorithms to register–transfer statements and datapaths*	8.5
12. *Designing control for datapaths*	8.5
13. *Debugging, troubleshooting, and designing simulations*	8.5
14. *Binary arithmetic*: Topics such as binary addition and subtraction, but not optimized circuits	8.4
15. *Using CAD tools*	8.4

Adapted from Goldman et al. (2010).

Based on the Delphi process, investigations to identify students' misconceptions focused on concepts concerning state, medium-scale integrated (MSI) circuits, Boolean expressions, and number representations. Topics such as *Using computer-aided design (CAD) tools* and *designing control for datapaths* were excluded because they either represented skills rather than concepts or were still highly contentious among the panel of experts (e.g. many institutions do not teach CAD tools) (Herman, Loui, & Zilles, 2010).

Students' misconceptions were identified using clinical interviews in which students verbalized their reasoning as they solved digital logic textbook problems. Since documenting the identified misconceptions is beyond the scope of this paper, interested readers are referred to other publications (Herman et al., 2011a, 2011b, 2012; Herman et al., 2012; Montfort et al., 2012). One key finding from the interviews was the context-dependence of students' conceptual knowledge: Students revealed specific misconceptions while solving certain types of problems, but did not reveal those same misconceptions when solving other problems or even an isomorphic problem with a different presentation style (Herman, Loui et al., 2012). This type of response pattern aligns well with DiSessa's KiP, which argues that novices lack coherent conceptual frameworks; instead, novices have collections of isolated facts that they cue or access based on surface-level features of a problem (diSessa, Gillespie, & Esterly, 2004). For example, when students

translate the English specification of "not both" into a Boolean expression, the phrase "not both" cues their use of the exclusive-or (XOR) concept rather than the NOT AND (NAND) concept (Herman, Loui et al., 2012). However, when the students translate the same English specification into a truth table, the presence of the truth table cues their conceptual knowledge of systematically testing all cases, and the students are able to recognize the "not both" specification as the NAND concept (Herman, Loui et al., 2012).

The KiP theory contrasts with the competing naive theories' understanding of conceptual knowledge, which posits that students develop naive or proto-theories that are coherent but limited in their explanatory power (Chi & Slotta, 1993; diSessa et al., 2004; Reiner, Slotta, Chi, & Resnick, 2000; Vosniadou & Brewer, 1992). For example, children possess a naive theory that there are two earths: a flat earth on which people stand and a spherical earth that hangs in space. This theory allows children to reason about some questions well (e.g. "Where is North America relative to Australia? Opposite side of a globe."), but fails in other contexts (e.g. "What will happen if I keep walking east? I will fall off.") (Vosniadou & Brewer, 1992). Concept inventories that rely on naive theories will attempt to discover what naive theories a student possesses. In contrast a concept inventory that relies on KiP will attempt to classify whether students possess sufficient coherence in their understanding to align with accepted disciplinary theories. Consequently, items for the DLCI focus on contexts that reveal coherence breaks in students' understanding.

2.2. The observation basis of the DLCI

Based on the cognition corner, we chose to create items (multiple-choice questions) that would yield observations based on two criteria: (1) items should test concepts associated with the important topics identified by the Delphi process and (2) items should test for the presence of students' misconceptions as revealed by the clinical interviews (Herman et al., 2010). To satisfy the first criterion, all items had to be related to the general topic areas of number representations, Boolean logic, MSI, and state. The Delphi process allowed us to choose this subset of core topics to observe. These topics were non-comprehensive, yet representative of the knowledge that students would be expected to know after a first course in digital logic.

Because the goal of a concept inventory is to reveal how and where students possess misconceptions and because items were constructed based on KiP and the clinical interviews, we constructed items only if the interviews had revealed contexts and situations that primed students to reveal misconceptions. Because the elicitation of students' misconceptions was often constrained to a single context, many concepts appear in only one item, so as to avoid testing students with isomorphic items that do not offer additional information about students' conceptual understanding. Adding items that

do not reveal students' misconceptions would be undesirable as these items would simply add to the length of the instrument without providing useful information about the limits of students' understanding. For example, no DLCI item asks students to translate the English specification "not both" into a truth table, because that context does not possess the contextual cue that prompts the misconception of "not both" meaning XOR.

A companion document to the DLCI was created to elaborate on which misconceptions are revealed by each *distractor* (a wrong multiple-choice answer). This document is available upon request.

The DLCI has currently iterated through four major versions as items were added or removed based on student performance and expert feedback. For example, items on implication (if-then logical statements) were removed after many experts stated that the concept was not taught or covered in their courses (Herman et al., 2010).

2.3. The interpretation basis of the DLCI

Because the DLCI was built on KiP and as a survey of core topics in digital logic, we assert that scores on the DLCI should be interpreted holistically: *a students' score on the DLCI should be treated as an estimation of that students' level of expertise*. This score reflects both the presence (or absence) of misconceptions and a student's ability to recognize which *schema* (organized patterns of thought) and conceptual chunks to use during a given context.

In order to verify our assertion, the CI must undergo validity and reliability testing. The content and face validity (i.e. does the concept inventory span the appropriate and desired technical domain? (Streveler et al., 2011)) has been previously established by the panel of Delphi experts who reviewed the content of the DLCI (Herman et al., 2010). The majority of experts indicated that the DLCI reflects core conceptual knowledge (Herman et al., 2010). The majority also indicated that they were confident that the DLCI would be a good predictor of student conceptual understanding in a first course on digital logic (Herman et al., 2010). Some experts expressed concerns that the DLCI does not assess problem-solving skills and is not comprehensive, but these two objectives are contrary to the objectives of a CI (Herman et al., 2010).

Construct validity is an evaluation of whether the items test the concepts that we intend to test (Streveler et al., 2011). The construct validity of the DLCI was also established by having the Delphi experts rate the quality of each question and suggest improvements. Additionally, construct validity should be established through rigorous analysis of students' performance on the DLCI. Students' selection of common misconceptions through the distractors has previously been verified through think-aloud protocols (Herman et al., 2010), in which students verbalize their thoughts as they solve items from the DLCI. These studies revealed that students

select distractors for the reasons that we predicted based on the clinical interviews (Herman et al., 2010).

3. Exploring the interpretation corner for the DLCI

The field of psychometrics provides a variety of statistical measures and techniques for evaluating the quality of an assessment instrument (Pellegrino et al., 2011). These measures inform how results on the instrument should be interpreted and to what types of decisions these interpretations should be applied (e.g. high-stakes admissions decisions vs. low-stakes formative feedback). The Learning Sciences Research Institute has outlined a set of best-practices for applying psychometrics to the validation and interpretation of concept inventories (DiBello et al., 2013; Jorion et al., 2014a; Pellegrino, DiBello, Miller, & Streveler, 2013). These practices have been applied to STEM concept inventories such as the Thermal and Transport Concept Inventory (TTCI) (Jorion et al., 2014b; Streveler et al., 2011), the Conceptual Assessment for Statics (CATS) (Jorion, James, Schroeder, & DiBello, 2013; Steif & Dantzler, 2005), and the Statistics Concept Inventory (SCI) (Jorion et al., 2013; Stone 2006).

Based on the framework laid out by Jorion et al. (2014a), we evaluated the DLCI according to principles from Classical Test Theory (CTT) and Item-Response Theory (IRT). If the instrument proves satisfactory according to CTT and IRT, then exploratory factor analysis and other clustering techniques may be used to determine whether subscales of the instrument can be used to measure students' mastery of specific concepts or topic areas (Jorion et al., 2013; Pellegrino et al., 2011). The following statistical analyses were performed on version 4.0 of the DLCI. We provide comparative data from the TTCI, CATS, and SCI where appropriate to aid in the interpretation of the results from the DLCI, because these concept inventories have been analyzed by a common framework.

3.1. Classical test theory

CTT is one branch of psychometrics for evaluating assessment instruments. CTT is primarily concerned with the reliability and validity of the test as a whole, but also provides some insights into the quality of each item. An ideal instrument should provide a reliable and valid measurement of each student's ability with a minimal number of items to span the requisite topic (Streveler et al., 2011).

3.1.1. Reliability

According to CTT, an assessment tool estimates a student's *true score* (T) as a function of the student's *actual score* (X) and some error (E) in the

Table 2. Cronbach alpha (α) calculated for the whole DLCI version 4.0, after it was administered at six institutions across the United States.

DLCI version	α	Sample size
4.0 (pre-test)	0.54	377
4.0 (post-test)	0.80	900

measurement ($T = X + E$) (Lord & Novick, 1968). A reliable test minimizes the expected error of measurement, so that if a student were to take the same instrument multiple times, it would yield the same measurement consistently (Kuder & Richardson, 1937). Cronbach's alpha (α) is a common measure of reliability, because it does not require that each examinee take the instrument multiple times (Bechger, Maris, Verstralen, & Beguin, 2003; Wallace & Bailey, 2010). Cronbach's α may be when two conditions are met: 1) the instrument measures a single latent trait and (2) items must be scored dichotomously (0 or 1) as correct or incorrect. Reliable instruments will yield α-values that are close to one (Borsboom, 2005). Some sources consider instruments as reliable for research if α is 0.80 or above (Bechger et al., 2003; George & Mallery, 2009; Jorion et al., 2013), whereas others consider instruments with α of 0.70 or above as acceptable (Nunnally, 1978; Stone, 2006). A Cronbach α of 0.60 or above is generally considered acceptable for typical classroom assessments (Bechger et al., 2003; George & Mallery, 2009; Jorion et al., 2013; Nunnally, 1978).

The standard error of measurement ($SE = S_x\sqrt{1-\alpha}$) is a function of the reliability of the instrument (α) and the standard deviation of the sample S_x (Lord & Novick, 1968). The standard error of measurement can be used to provide a confidence interval for each examinee's true score.

The DLCI Version 4.0 (or β1.0 as referenced in other publications (Herman et al., 2010; Herman & Loui, 2011) has 24 items. For the research reported here, the DLCI has been administered as a post-test to exactly 900 students distributed across six institutions across the United States: three large public research universities (one each in the Midwest, on the West coast, and on the East coast) and three small private colleges (one in the Midwest and two in the South). This sample has yielded Cronbach's α of 0.80, mean of 14.53, standard deviation of 4.76, and standard error of measurement of 1.89. The α values of the TTCI, CATS, and SCI are 0.67 (11 items), 0.84 (27 items), and 0.64 (37 items), respectively, when administered as post-tests (Jorion et al., 2013).

The DLCI has also been administered as a pre-test to a sub-population of 377 students from the 900 students. This sample yielded an α of 0.54, mean of 7.93, standard deviation of 3.26, and standard error of measurement of 2.22 (Table 2).

Table 3. Cronbach α calculated excluding individual items from the DLCI. Each column indicates which item was excluded for each calculation.

Item excluded	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Q7*	*Q8*	*Q9*
Cronbach's α	0.79	0.79	0.79	0.80	0.79	0.79	0.79	0.80	0.79
Item excluded	*Q10*	*Q11*	*Q12*	*Q13*	*Q14*	*Q15*	*Q16*	*Q17*	*Q18*
Cronbach's α	0.79	0.78	0.79	0.79	0.79	0.80	0.79	0.79	0.79
Item excluded	*Q19*	*Q20*	*Q21*				*Q22*	*Q23*	*Q24*
Cronbach's α	0.79	0.78	0.79				0.79	0.79	0.79

Table 4. Cronbach α calculated for each sub-topic area of the DLCI.

Subtopic	Cronbach's α	Items included
Number representations	0.57	6 items (*Q3*, *Q10*, *Q14*, *Q16*, *Q22*, *Q23*)
Boolean logic	0.57	7 items (*Q1*, *Q5*, *Q8*, *Q9*, *Q19*, *Q20*, *Q21*)
MSI	0.45	5 items (*Q4*, *Q11*, *Q12*, *Q13*, *Q24*)
State	0.56	6 items (*Q2*, *Q6*, *Q7*, *Q15*, *Q17*, *Q18*)

Cronbach's α can also provide coarse information about item quality (Bock & Lieberman, 1970; Jorion et al., 2013). Adding an item should increase the reliability of the instrument (Bock & Lieberman, 1970; Jorion et al., 2013). Items that decrease the reliability should be removed from the instrument or inspected for errors. Table 3 shows that excluding any one item causes the α to decrease (excluding one of items *Q4*, *Q8*, or *Q15* lowers the Cronbach α, but not noticeably at the chosen number of significant figures), indicating that each of the items contributes to improving the reliability of the DLCI.

Cronbach's α can also be applied to subscales of the instrument to test whether subsets of items can be used to estimate students' ability on a sub-skill (Pellegrino et al., 2011). Items on the DLCI were categorized according to the sub-topics indicated by the Delphi process: Number representations, Boolean logic, MSI, and state. Table 4 shows the Cronbach αs as calculated for the subscales of the DLCI as indicated by the Delphi process. None of the subscale α values is above the desired level of 0.60. The TTCI has three subscales with α values of 0.52, 0.54, and 0.59 (Jorion et al., 2013). The CATS has nine subscales with αs ranging from 0.33 to 0.72 and only three subscales below 0.60 (Jorion et al., 2013). The SCI has four subscales with α values of 0.47, 0.27, 0.43, and 0.39 (Jorion et al., 2013).

3.1.2. Validity

Reliability is a necessary, but not sufficient condition for the validity of the instrument (i.e. if an instrument is unreliable, it cannot provide a valid measurement of ability) (Pellegrino et al., 2013). Aside from content and

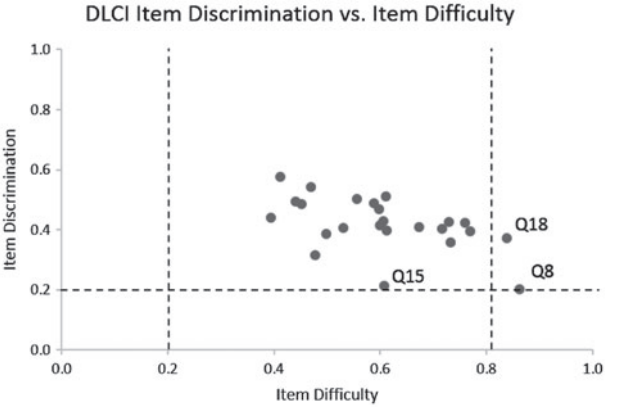


Figure 3. Plot of item discrimination vs. difficulty of the DLCI.

construct validity (discussed in Section 2.3), the validity of an instrument can be further established by the statistical properties of the instrument with respect to difficulty and discrimination.

In CTT, item difficulty is the fraction of examinees that answered an item correctly (0.0 would be an impossible item, 1.0 would be a trivial item) (Crocker & Algina, 1986; Lord and Novick, 1968). A valid instrument will have items than span the range from easy to difficult. *Item discrimination* is the Pearson (point-biserial) correlation coefficient between examinees' performance on each item and their performance on the instrument as a whole. An item with greater discrimination provides more information about the student's ability level. Items should have greater than 0.2 discrimination (horizontal dotted line in Figure 3) and between 0.2 and 0.8 item difficulty (vertical dotted lines in Figure 3) (Bardar, Prather, Brecher, & Slater, 2007; Ding & Beichner, 2009; Ding, Chabay, Sherwood, & Beichner, 2006; Jorion et al., 2013; Kline, 2005). An instrument may have up to three exceptions and still be considered a valid instrument (Jorion et al., 2013). Items that fail to meet these criteria, may be kept given sufficient rationale through other metrics (such as reliability or construct validity).

The DLCI reveals the desired instrument properties, although items *Q8*, *Q15*, and *Q18* merit additional scrutiny to determine whether they should be improved or removed from the DLCI as they possess weak discrimination or may be too easy. Because CTT is primarily concerned with properties of the instrument as whole, we use IRT to better understand the properties of each item individually.

3.2. Item-response theory

While CTT evaluates the quality of items within the context of a specific instrument and specific population, IRT generalizes the performance of each

item, allowing each item to be treated as an independent item (Hambleton & Jones, 1993; Hambleton, Swaminathan, & Rogers, 1991). This property allows items to be reused and reconfigured into new instruments. Further, CTT assumes that the error of measurement is the same for every examinee while IRT assumes that the error of measurement can vary between examinees (i.e. measurement of an expert's ability may have less error than the measurement of a true novice's ability due to guessing) (Hambleton & Jones, 1993). Item response theory posits that the probability that an examinee will answer an item correctly ($p_i(\theta)$) is a mathematical function of the examinee's ability and item parameters. An examinee's latent trait ability level (θ) represents their ability level in a specific domain.

IRT specifies that θ has a mean of 0.0 and standard deviation of 1.0. A one-parameter logistic model (or Rasch model) constrains all items to have the same level of discrimination a_i, but items can vary according to difficulty (b_i) (Hambleton et al., 1991). Critically, item discrimination and difficulty in IRT are not equivalent to item discrimination and difficulty in CTT. In a two-parameter logistic model, each item in an instrument can vary in its difficulty (b_i) and discrimination (a_i) (see Equation (1)) (Hambleton et al., 1991).

$$p_i(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}} \tag{1}$$

where p_i is the probability that an examinee answers item i correctly, θ is an examinee's ability level, a_i is item i's discrimination parameter, and b_i is item i's difficulty parameter.

Item Response Functions (IRFs) provide a mathematical/graphical model for estimating the probability (y-axis) that an examinee with a given θ (x-axis) will answer an item with parameters a_i and b_i correctly. The IRF of a good item will asymptotically approach 1 as θ increases and asymptotically approach 0 as θ decreases (see Figure 4). Item difficulty b_i is the θ at which the IRF crosses 0.50, indicating that a person with ability level θ has a 50% chance of answering that item correctly. The item discrimination determines the slope with which the curve crosses through 0.50. A steeper slope is more desirable as it offers greater power to distinguish between examinees of similar θ (Hambleton et al., 1991).

An *Item Information Curve* (IIC) measures how much information an item provides about an examinee with a given θ (Hambleton et al., 1991). The information function for an item is defined in Equation (2), where a_i is the item discrimination parameter, $p_i(\theta)$ is the probability that the examinee would get the item correct, and $q_i(\theta)$ is the probability that the examinee would get the item wrong (i.e. $1 - p_i(\theta)$) (Hambleton et al., 1991).

$$I_i(\theta) = a_i^2 p_i(\theta) q_i(\theta) \tag{2}$$

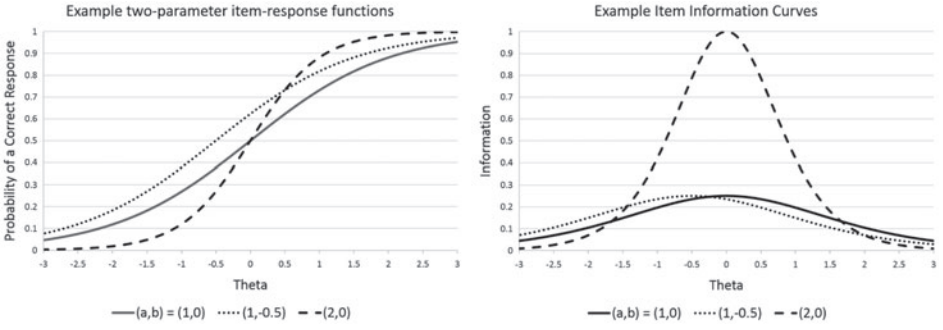


Figure 4. Left: Three example item response functions with varying discrimination (a_i) and difficulty (b_i). Right: Item information curves for the same example items.

The bell curve shape of the item information curve indicates that an item provides the most information about examinees whose θ is close to the item's difficulty. Because items are considered to be independent, the amount of information that an instrument provides about an examinee of ability θ is the summation of the IICs of all items. The summation of information for an entire instrument dictates how an instrument should be used. If the goal of an instrument is to distinguish between examinees who pass or fail based on a single cutoff, an instrument should have information clustered at the cutoff point (Hambleton et al., 1991). Instruments that are intended to distinguish between a range of abilities should provide information across the spectrum of ability levels.

The *test information curve* is the summation of all item information curves (Hambleton et al., 1991). In Equation (3), n is the number of items in the instrument.

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \tag{3}$$

The standard error of measurement function for an instrument is defined as

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}. \tag{4}$$

Critically, according to IRT, the accuracy with which an instrument can estimate an examinee's θ depends on that examinee's θ (Hambleton et al., 1991). For example, if an instrument is constructed to provide a lot of information about examinees with low ability levels, it will not provide an accurate estimate of an examinee's ability should they have a high ability level. This assumption stands in contrast to CTT which assumes that the standard error of measurement is the same for all examinees (inversely proportional to Cronbach's α).

3.2.1. Estimation of ability and item parameters

An examinee's ability θ is estimated not by the total score as in CTT, but according to the probability that the examinee achieved a given response pattern u (Hambleton et al., 1991). The likelihood function for an examinee response pattern is given in Equation (5), where n is the number of items on the instrument, u_i is the examinee's response to item i, P_i is the probability of a correct response to item i, and Q_i is the probability of an incorrect response. For example, for a dichotomously scored, three-item instrument, an examinee might have a response pattern $u_1 = 0$ (incorrect), $u_2 = 1$ (correct), and $u_3 = 1$ (correct), yielding a likelihood function $L(u_1, u_2, u_3|\theta) = Q_1 P_2 P_3$. The value of θ that maximizes Equation (5) is defined as the maximum likelihood estimate of that examinee's ability level.

$$L(u_1 \ldots u_n|\theta) = \prod_{i=i}^{n} P_i^{u_i} Q_i^{1-u_i} \tag{5}$$

Because item parameters and examinee ability levels are both unknown for the DLCI, they must be jointly estimated based on the data. This estimation is performed using a joint maximum likelihood estimation defined in Equation (6), where N is the number of examinees in the data set, $u_{1j}, \ldots, u_{nj}$ is the response pattern for examinee j, P_{ij} is the probability that examinee j chooses a correct response to item i, Q_{ij} is the probability of an incorrect response, a is a vector of candidate item discriminations, and b is a vector of candidate item difficulties (Hambleton et al., 1991).

$$L(u_{1,1}, u_{1,2}, \ldots, u_{n,N}|\theta, a, b) = \prod_{i=1}^{n} \prod_{j=1}^{N} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \tag{6}$$

The maximum likelihood estimate cannot be computed, unless we can assume that the sample population has a normal distribution with a mean θ of 0 and standard deviation of 1. The sample must also possess a sufficient number of examinees from across the distribution of abilities to provide reliable parameter estimation (Hambleton et al., 1991). Each additional parameter to be estimated requires estimation over one additional dimension. Consequently, the one-parameter Rasch model is often preferred to the two-parameter model for computational simplicity and has been the predominant model for the evaluation of other concept inventories (Barbera, 2013; Libarkin & Anderson, 2006; Wren & Barbera, 2014). Item parameters and ability estimates for the DLCI were calculated according to the marginal maximum likelihood estimate method as developed by Bock and Lieberman (1970).

Once parameters and ability levels are estimated for a set of items, they can be evaluated for their goodness of fit using Pearson's χ^2 statistic (Rizopoulos, 2006). The null hypothesis of the χ^2 test is that the sample

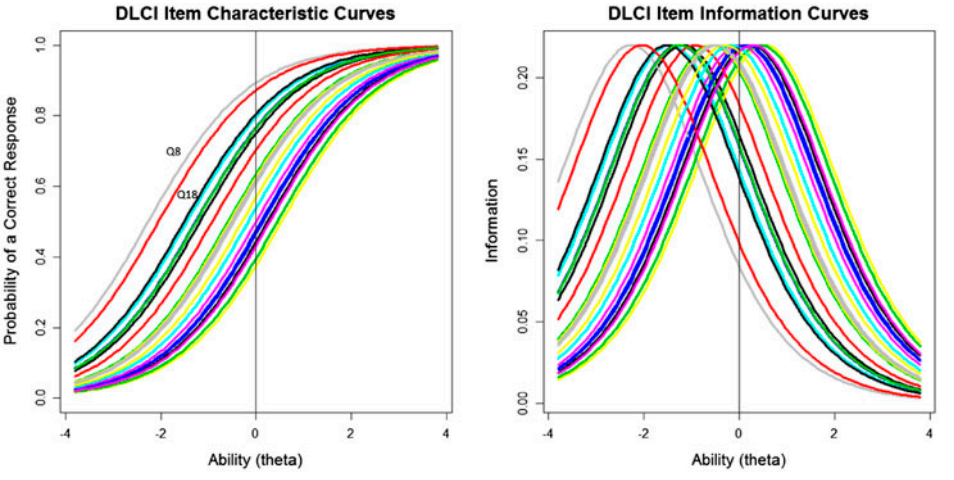


Figure 5. Result from the single parameter Rasch model. Left: Item response functions for all items on the DLCI with the constraint $a_i = 0.94$. Items *Q8* and *Q18* are shown again to be the easiest items. Right: Item information curves for the same items.

distribution and the ideal distribution derived from the model parameters are the same distribution. Consequently, a higher p-value indicates a higher probability that the sample distribution and ideal distribution are the same. Therefore, non-significant p-values indicate an acceptable fit of the data to the model.

3.2.2. Application of IRT to the DLCI

Because of the common practice of using the Rasch model for other concept inventories, items on the DLCI were modeled with IRF curves based first on the single parameter Rasch model (a_i (discrimination) is constrained to be the same for all items, b_i (difficulty) is variable across items). We also explored the two-parameter model (a_i and b_i are both variable) to examine whether it provided a better model for our data.

IRFs and IICs for the single parameter Rasch model are shown in Figure 5. The goodness of fit test yielded a p-value of 0.56, indicating an acceptable fit of the data to the model. Item difficulty parameters range from two standard deviations below the mean ($b_8 = -2.26, b_{15} = -2.04$) to half of a standard deviation above the mean ($b_7 = 0.55$).

IRFs and IICs for the two-parameter model are shown in Figure 6. We used a Likelihood Ratio Table to compare the goodness of fit between the two models. The null hypothesis for this test is that the two models have the same quality of fit. The test revealed that the two parameter model provided a better fit for the data ($p < 0.01$). Each item's difficulty and discrimination parameters can be found in Table 5. The variance in discrimination parameters as seen in Table 5 confirms that constraining all

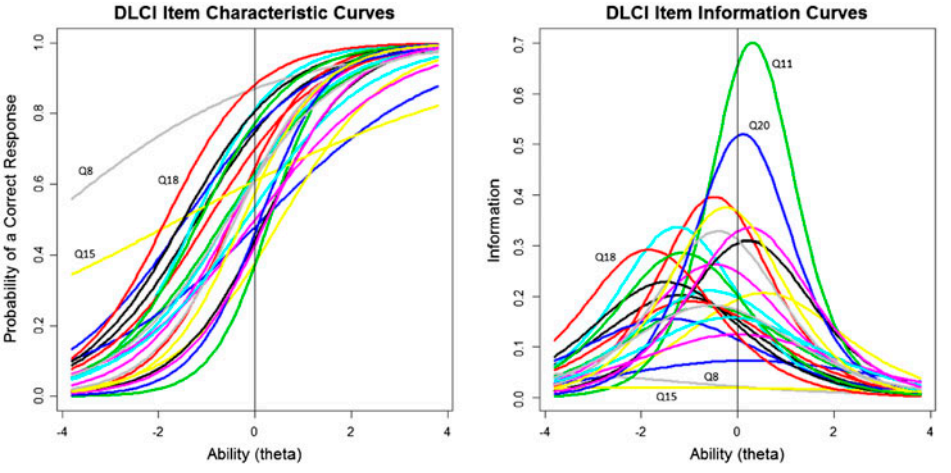


Figure 6. Result from the two-parameter IRT model. Left: Item response functions for all items on the DLCI unconstrained difficulty and discrimination parameters. Right: Item information curves for the same items. Items *Q8* and *Q15* are shown to provide little information about examinees, whereas item *Q18* provides adequate information.

items to have the same discrimination (as required in the Rasch model) may be a poor assumption. Items *Q8* and *Q15* have low discrimination values ($a_i < 0.50$) and offer little information about examinees of any ability level (see Figure 6). Item *Q8* is also considerably easier according to this model ($b_8 = -4.32$) as compared to the Rasch model. Item *Q18* is more difficult according to this model ($b_{18} = -1.84$) as compared to the Rasch model. Items *Q11* and *Q20* provide the most information, particularly for examinees with θ slightly above the mean (($a_{11}, b_{11}) = (1.68, 0.31)$, ($a_{20} = 1.44, b_{20} = 0.11$)).

The DLCI test information curve reveals that the DLCI provides maximum information (5.4) for students with $\theta = -0.25$ (see Figure 7). The test information curve remains above 4 for $-1.6 < \theta < 0.9$, yielding standard errors of measures below 0.5 over that range of ability levels. Consequently, the DLCI can estimate a students' ability level θ within ± 0.5 standard deviations with 68% confidence.

4. Discussion

The psychometric analysis of the DLCI builds on the previous work of Jorion et al. (2013) in establishing a framework for rigorously analyzing concept inventories. The goal of this discussion is to provide an example of how to interpret psychometric analysis of concept inventories for other computer science education researchers. We first discuss the quality of the items on the DLCI and then discuss the quality of the DLCI as a whole.

Table 5. Item difficulty (b_i) and discrimination (a_i) for all DLCI items as determined by the two-parameter IRT model.

Item	Difficulty	Discrimination	Item	Difficulty	Discrimination
Q1	−1.21	0.90	*Q13*	−0.56	0.92
Q2	−0.48	1.26	*Q14*	0.01	0.71
Q3	−0.63	0.85	*Q15*	−1.55	0.28
Q4	0.17	0.54	*Q16*	−0.41	1.15
Q5	−1.25	1.16	*Q17*	−1.49	0.96
Q6	−0.48	1.03	*Q18*	−1.84	1.08
Q7	0.55	0.91	*Q19*	−1.14	1.07
Q8	−4.32	0.44	*Q20*	0.11	1.44
Q9	0.21	1.11	*Q21*	−0.18	0.80
Q10	−0.97	0.87	*Q22*	0.25	1.16
Q11	0.31	1.68	*Q23*	−0.25	1.22
Q12	−1.44	0.79	*Q24*	−0.55	0.85

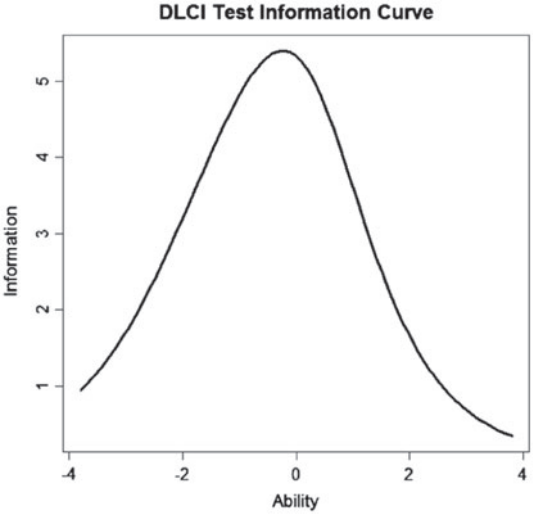


Figure 7. Test information curve for the DLCI.

Discussion will focus on the results of CTT and the two-parameter model from IRT.

4.1. Item analysis

In general, the DLCI exhibits desirable psychometric properties according to CTT. None of the items are detrimental to the reliability of the DLCI. Similarly, most items reveal desirable levels of difficulty (between 0.2 and 0.8) and discrimination (above 0.2). According to IRT, every item of the DLCI fits the two-parameter model. The DLCI also possesses a desired level of

information about students' conceptual understanding across a range of ability levels.

Only items *Q8*, *Q15*, and *Q18* were flagged as candidate items for improvement or removal from the DLCI by either CTT or IRT. Decisions to keep, improve, or remove items should consider the psychometric qualities of the item as well as the cognition and observation qualities of the items.

4.1.1. Analysis of item Q8

Item *Q8* (see Figure 8) was flagged for low discriminatory power by both CTT and IRT. It was also flagged as being too easy by both theories. According to IRT, item *Q8* is challenging only for students more than four standard deviations below the mean in ability ($b_8 = -4.32$).

Item *Q8* (correct answer is option 2) assesses whether students forget to include negated variables in their Boolean expressions (option 1) and initially had only options 1 and 2. Options 3 and 4 were added simply to provide four choices as an attempt to improve the item and are chosen less than 1% of the time. The item reveals that at least 15% of students possess the misconception of forgetting to include negated variables for such a simple expression. This misconception stems from a schema in which students fail to fully enumerate cases when constructing Boolean expressions. This schema is the basis for the misconceptions targeted by items *Q9* and *Q20*. Items *Q9* and *Q20* provide strong discrimination and are fairly challenging ($(a_9, b_9) = (1.11, 0.21)$, $(a_{20}, b_{20}) = (1.44, 0.11)$), demonstrating the prevalence and detrimental nature of this misconception. As described by Knowledge-in-Pieces, item *Q8* does not seem to provide a context that is appropriate for eliciting this common misconception. In particular, we suspect that because the correct response (option 2) provides the visual information to include the negated variables of *l* and *t* immediately below the misconception, it provides just enough contextual information to remind students about the need to include the negated variables.

Because previous attempts to improve item *Q8* have failed and because our cognitive theory suggests that this item may be fundamentally flawed in the multiple-choice context, this item will be removed from future versions of the DLCI.

4.1.2. Analysis of item Q15

Item *Q15* was flagged for low discriminatory power by both CTT and IRT. Neither CTT nor IRT flagged item *Q15* for its difficulty.

Item *Q15* (see Figure 9) assesses whether students conceive of state as being comprised of the inputs (x) and outputs (z) of the system (incorrect) or as just the information stored in elements such as flip-flops (correct – choice 3). Conceptually, the input x can influence the state of circuit G

For Questions 8 and 9, suppose that a sandwich shop has only the following ingredients.

b = bacon, l = lettuce, t = tomato

Question 8. Alice requires that a sandwich must have bacon by itself. Which Boolean expression correctly specifies all sandwiches that satisfy her requirement?

1) $\boldsymbol{Alice} = b$
2) $\boldsymbol{Alice} = b\,\bar{l}\,\bar{t}$
3) $\boldsymbol{Alice} = \bar{l} + \bar{t}$
4) $\boldsymbol{Alice} = \bar{l}\,\bar{t}$

Figure 8. Item *Q8* from the DLCI assesses whether students forget to include negated variables in their Boolean expressions.

only when the *clk* signal rises from 0 to 1. When reasoning about timing diagrams, students indiscriminately fixate on times when any signal changes its value and declare that the state of the system has changed. The incorrect choices in item *Q15* use dotted ovals to highlight moments where x, z, and *clk* change their values, mimicking student behavior during think-aloud interviews. This notation is a new notation that students have never seen before.

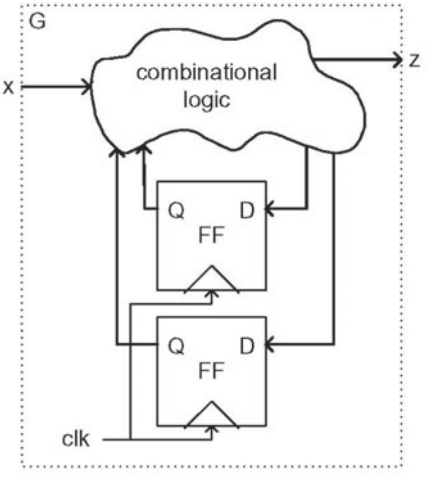
We suspect that the item's poor discriminatory power stems from the use of this new notation. This item seems to test two aspects of students' knowledge – presence of misconceptions and ability to interpret the new notation – rather than just whether students possess misconceptions about what comprises state in a circuit. The presence of this misconception is also assessed in items *Q6* ($(a_6, b_6) = (1.03, -0.48)$) and *Q17* ($(a_{17}, b_{17}) = (0.96, -1.49)$), both of which have acceptable discriminatory power. Item *Q15* could potentially be improved with a better notation. If not, it will be removed from the DLCI. Future improvement of this item could be informed by performing additional think-aloud interviews with students as they solve this item.

4.1.3. Analysis of item Q18

Item *Q18* has an acceptable level of discriminatory power according to both CTT and IRT ($a_{18} = 1.08$). CTT flagged Item *Q18* as being too easy. IRT agrees that item *Q18* is easy ($b_{18} = -1.84$), but its difficulty level is not alarmingly low as it is still challenging for students with ability between one and two standard deviations below the mean.

Item *Q18* (see Figure 10) assesses whether students struggle to distinguish between conceiving of the state of a circuit as a whole versus the state stored in individual sub-components of a circuit as well as whether they conceive of state being stored in the inputs or outputs of the circuit. These misconceptions are particularly detrimental as they inhibit students' ability

Question 15.



The block diagram of a synchronous finite state machine – circuit G – is shown above. Circuit G is composed of combinational logic (of unknown design) and positive edge triggered D flip-flops. Circuit G also has one input *x* and one output *z*. Suppose the timing diagrams below are produced by circuit G. **Which timing diagram correctly indicates the times when *x* influences the state of circuit G**?

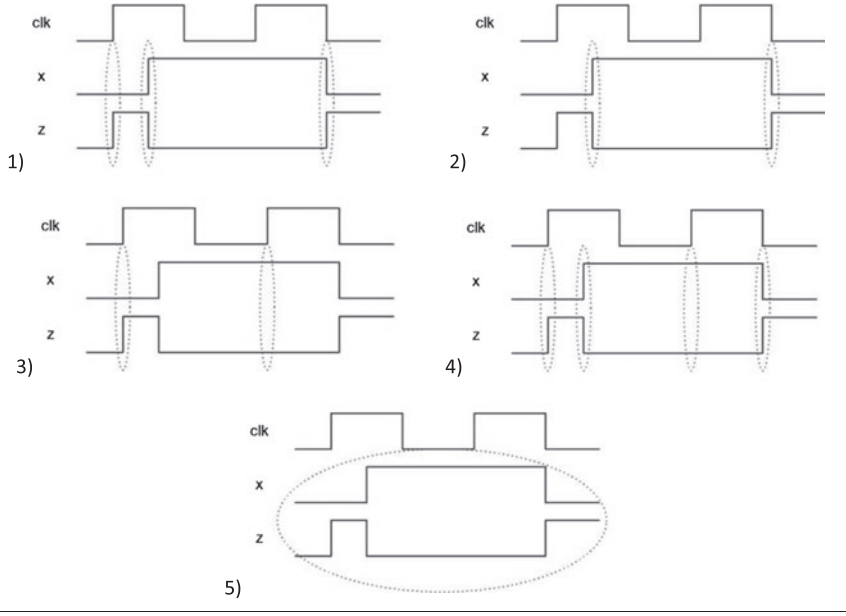


Figure 9. Item *Q15* from the DLCI assesses whether students mistakenly believe that the state of the circuit is determined by the values of its inputs (x) and outputs (z).

to interpret or create finite state machines. It is also the only item on the DLCI that assesses students' ability to distinguish between the state of a component and the state of a circuit. We believe that the item can be kept

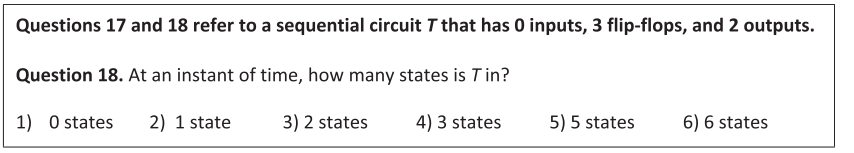


Figure 10. Item *Q18* from the DLCI assesses whether students mistakenly believe that a circuit can exist in multiple states at a time.

in the DLCI. Expert feedback on the item has been positive, and observing students as they answer the item has not revealed any structural concerns with the item.

4.2. Appropriate uses of the DLCI

The CTT analysis of the DLCI provides psychometric evidence that the DLCI can be used as a reliable and valid post-test instrument for assessing students' conceptual understanding of digital logic. As shown in Section 3.1.1, the DLCI has a high reliability coefficient according to CTT. The DLCI is as reliable or more so than other similarly evaluated concept inventories (Jorion et al., 2013). It also provides an acceptably low standard error of measurement for students across a wide range of ability levels ($SE(\theta) < 0.75$ for examinees with $-3 < \theta < 2$). The DLCI can reliably and validly classify students as having strong versus weak conceptual understanding of digital logic concepts, supplying evidence for discriminant validity. Consequently, the DLCI may be used for comparing the effectiveness of instructional methods for digital logic or providing estimates for the ability level of students.

The evidence does not support using the DLCI as a reliable and valid pre-test for assessing students' conceptual understanding. The pre-test administrations of the DLCI have unacceptable reliability and larger standard errors than post-test administrations. This result is expected. The validation attempts reported for other concept inventories report similar unreliability of pre-test administrations (Steif & Dantzler, 2005; Wallace & Bailey, 2010). Like other computing topics, digital logic relies on terminology and notation that students are unlikely to have encountered prior to instruction, such as multiplexers and flip-flops, so most students are likely guessing on many, if not most, items (Almstrum et al., 2006; Porter, Garcia, Tseng, & Zingaro, 2013; Taylor et al., in press; Tew, 2010; Webb & Taylor, 2014).

Accordingly, the DLCI should not be used to measure the normalized gain as proposed by Hake (1998). The normalized gain compares a population's post-test scores with their pre-test scores to measure what percentage of information the population learned that they did not already know. Although the normalized gain metric has commonly been used in studies that use concept inventories, the validity of the metric has been increasingly

called into question by other studies (Wallace & Bailey, 2010). The only valid use of the DLCI is as a post-test that assesses how much students' conceptual understanding aligns with accepted conceptual frameworks after instruction.

The DLCI appears to measure a single construct: conceptual understanding of digital logic. IRT assumes a unidimensional latent trait θ, and the DLCI demonstrates a good fit to this assumption. Similarly, the DLCI does not possess sufficient reliability within expert identified subscales of the DLCI. The subscales should not yet be used independently in high-stakes contexts such as research. The reliability metrics are borderline for classroom assessment, so subsets of items may be useful for formative feedback such as clicker questions (classroom response systems for large lecture courses) and other low-stakes contexts. Alternatively, subsets of items could be used in higher stakes classroom assessments if supplemented with additional items.

4.3. *Improving the DLCI*

Based on the item analysis, we removed items *Q8* and *Q15* from the DLCI to create version 4.1. Removing the items maintained the reliability of the DLCI at Cronbach $\alpha = 0.80$, indicating that the removal of these items is acceptable according to CTT. Removing items *Q8* and *Q15* improved the fit of the Rasch model to the data (χ^2 statistic decreased yielding a higher p-value, $p = 0.74$). The two-parameter model still proved to be a better fit for the data than the Rasch model ($p < 0.01$). Removing the items did not noticeably reduce the DLCI information curve, maintaining the range of acceptable standard error of measurement.

With only 22 items, the DLCI is shorter than other STEM concept inventories, which often contain between 25 and 35 items (Steif & Dantzler, 2005; Stone, 2006; Wallace & Bailey, 2010). More items could be added to future versions of the DLCI without making the instrument onerous for students and instructors.

4.4. *Implications of the psychometric analysis for research on student cognition and learning of digital logic*

Interpreting the results of the DLCI should provide additional insights and research questions for the cognition and observation corners of the assessment triangle (Figure 1). In this subsection, we explore future directions for research inspired by the psychometric analysis.

Because the psychometric analysis failed to reveal coherent subscales, the analysis raises some questions about the conceptual structure of the DLCI. In particular, the categories that we identified during the Delphi process of number representations, Boolean logic, MSI, and state may be

topic categories rather than concepts. For example, we have previously proposed that the concept of organizing bits into groups to create meaning is a cross-cutting concept that appears in each of the existing topic categories: Groups of flip-flops create state, groups of bits create number representations, groups of select bits determine which data input is connected to the multiplexer output, and so forth. Interestingly, five of the ten most discriminating items from the DLCI v4.0 appear to be testing this concept (Items *Q2*, *Q11*, *Q16*, *Q22*, *Q23*). This observation suggests a hypothesis that the ability to consistently and reliably aggregate bits into different types of meaning across contexts is a primary distinction between novices and experts.

To test this hypothesis, we need to perform a domain analysis (Leech & Onwuegbuzie, 2007) to understand the interconnections between concepts and develop a rigorous model of the conceptual structure of digital logic. Additionally, for revisions to the DLCI, we could create new items that test this concept, which could produce additional evidence for testing this hypothesis.

Our examination of which items provided the greatest discrimination also revealed that four of the ten most discriminating items from the DLCI v4.0 (items *Q5*, *Q9*, *Q19*, *Q20*) involved misconceptions that stem from a schema (an organized pattern of thought) that leads students to incompletely enumerate test cases (*proof by incomplete enumeration* (Herman, Loui et al., 2012)), particularly with respect to the XOR concept. This schema was documented as we interviewed students and was a substantial piece of evidence for initially choosing the Knowledge-in-Pieces (KiP) framework. During interviews, students with robust conceptual knowledge revealed a variety of schemas appropriate to the different contexts. In contrast, students with weak conceptual knowledge relied on generating and evaluating incomplete sets of test cases before *satisficing* (finding a sufficiently satisfying or sufficing example) (Manktelow, 2000). Because we constructed items based on features that provoked students to use proof by incomplete enumeration, these items provide evidence that justifies our use of the Knowledge-in-Pieces framework and supports our qualitative findings that proof by incomplete enumeration is a detrimental schema that hinders novices (Herman, Loui et al., 2012).

The psychometric analysis of the DLCI provides evidence that justifies our use of a KiP framework for creating the DLCI. The lack of strong correlation between items that test similar concepts suggests variance in ability across contexts. Students' performance on item *Q8* stands in contrast to our observations during open-ended interviews in which students generated Boolean expressions, further suggesting the context dependence and fragmentation of students' knowledge. During interviews, students routinely omitted negated variables as seen in item *Q8* (Figure 8), but only the weakest students struggle with that error when the negated variables

were present just a few millimeters away among the possible answers on the printed test. However, when negated variables are omitted in other more subtle items such as items *Q1* or *Q20*, students revealed this common misconception with greater frequency than in item *Q8*.

The context-dependence of students' knowledge in the multiple-choice environment of the DLCI suggests future research on the effect of reordering items or answer choices. Would changing the ordering of items or their choices change the psychometric properties of the instrument? Might some orderings minimize the appearance of certain misconceptions or improper schema?

Because the DLCI is based on a KiP framework and does not possess reliable subscale instruments, the DLCI's reliability and information functions rely on assessing students' knowledge across a variety of contexts to provide useful information. It may be more accurate to claim that the DLCI provides an estimate for the robustness of a student's ability to identify and use digital logic concepts across a variety of contexts. Performing a domain analysis and revising the DLCI will enable deeper investigations into how to best assess students' understanding of digital logic and whether the conceptual framework identified by the Delphi process or the learning theory best explain student performance on the DLCI.

5. Conclusions

The psychometric analysis of the DLCI according to CTT and IRT extends prior validation efforts of the DLCI. We recommend the use of the DLCI in research and assessment of student learning only when used as a post-test and in its entirety. Sub-scales of the DLCI may be used as formative feedback for students, but should not be used as high-stakes assessment tools unless further supplemented. Finally, analysis of the DLCI supports prior findings about the origins and nature of students' misconceptions in digital logic, particularly the context-dependent nature of students' use of concepts and schemas. Future research should explore the conceptual structure of digital logic and the interplay between students' schemas and misconceptions.

Acknowledgements

Thanks to the Delphi experts for their input and feedback on the creation of the DLCI. Thanks to the instructors who have administered the DLCI and contributed their results to our psychometric analysis. A special thanks to James W. Pellegrino for providing initial support in sharing the methods used in this analysis.

Funding

This work was supported by the National Science Foundation under [grant number DUE-0618589] and [grant number DUE-1140554]. The opinions, findings, and conclusions do not necessarily reflect the views of the National Science Foundation or the authors' institution.

References

Almstrum, V. L., Henderson, P. B., Harvey, V., Marion, C. H. H., Riedesel, C., & Tew, A. E. (2006). Concept inventories in computer science for the topic discrete mathematics. *ACM SIGCSE Bulletin, 38*, 132–145.

Barbera, J. (2013). A psychometric analysis of the chemical concepts inventory. *Journal of Chemical Education, 90*, 546–553.

Bardar, E. M., Prather, E. E., Brecher, K., & Slater, T. F. (2007). Development and validation of the Light and Spectroscopy Concept Inventory. *Astronomy, 5*, 103–113.

Bechger, T. M., Maris, G., Verstralen, H., & Beguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27*, 319–334.

Bock, R., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179–197.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. New York, NY: Cambridge University Press.

Chi, M. T. H. (2006). Methods to assess the representations of experts' and novices' knowledge. In K. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 167–184). Cambridge: Cambridge University Press.

Chi, M. T. H., & Slotta, J. D. (1993). The ontological coherence of intuitive physics. Commentary on A. diSessas "Toward an epistemology of physics". *Cognition and Instruction, 10*, 249–260.

Clayton, M. J. (1997). Delphi: A technique to harness expert opinion for critical decision-making tasks in education. *Educational Psychology, 17*, 373–386.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovitch.

DiBello, L. V., Pellegrino, J., Miller, R., Streveler, R., Jorion, N., James, K., ... Stout, W. (2013). An analytical framework for investigating cis. In *Proceedings of the 2013 Annual Meeting of the American Educational Research Association*, San Francisco, CA.

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physics Re, 5*, 020103.

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief Electricity and Magnetism Assessment. *Physics Review Special Topics: Physics Education Research, 2*, 010105.

diSessa, A., Gillespie, N., & Esterly, J. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science, 28*, 843–900.

Evans, D. L., Gray, G. L., Krause, S., Martin, J., Midkiff, C., Notaros, B. M., ... Wage, K. (2003). Progress on concept inventory assessment tools. In *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference* (pp. T4G-1–T4G-8), Boulder, CO.

George, D., & Mallery, P. (2009). *SPSS for windows step by step: A simple guide and reference*. Boston, MA: Pearson Education.

Goldman, K., Gross, P., Heeren, C., Herman, G. L., Kaczmarczyk, L., Loui, M. C., & Zilles, C. (2010). Setting the scope of concept inventories for introductory computing subjects. *ACM Transactions on Computing Education, 10*, 5:1–29.

Hake, R. (1998). Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*, 64–74.

Hake, R. (2002). Lessons from the physics education reform effort. *Conservation Ecology, 5*, 28.

Hambleton, R. K., & Jones, R. J. (1993). Comparison of classical test theory and item response theory and their applications to test deveolpment. *Educational Measurement: Issues & Practice, 12*, 253–262.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage.

Herman, G. L. & Loui, M. (2011, June 26–29). Administering the digital logic concept inventory at multiple institutions. In *Proceedings of the 2011 American Society*

for Engineering Education Annual Conference and Exposition (pp. AC2011-1800), Vancouver, Canada.

Herman, G. L. & Loui, M. C. (2012, June 10–13). Identifying the core conceptual framework of digital logic. In *Proceedings of the 2012 American Society for Engineering Education Annual Conference and Exposition* (pp. AC2012-4637), San Antonio, TX.

Herman, G. L., Loui, M. C., Kaczmarczyk, L., & Zilles, C. (2012). Discovering the what and why of students' difficulties in Boolean logic. *ACM Transactions on Computing Education, 12*, 3:1–28.

Herman, G. L., Loui, M. C. & Zilles, C. (2010, March 10–13). *Creating the digital logic concept inventory*. In *Proceedings of the forty-first ACM technical symposium on computer science education* (pp. 102–106), Milwaukee, WI.

Herman, G. L., Loui, M. C., & Zilles, C. (2011a). Students' misconceptions about medium-scale integrated circuits. *IEEE Transactions on Education, 54*, 637–645.

Herman, G. L., Zilles, C., & Loui, M. C. (2011b). How do students misunderstand number representations? *Computer Science Education, 23*, 289–312.

Herman, G. L., Zilles, C., & Loui, M. C. (2012). Flipops in students' conceptions of state. *IEEE Transactions on Education, 55*, 88–98.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*, 141–166.

Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (2014a). Conceptual and analytical frameworks for examining validity and utility of concept inventories. In *Proceedings of the 2014 Annual Meeting of the American Educational Research Association*, Philadelphia, PA.

Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (2014b). Quantitative analyses of student performance on concept inventory. In *Proceedings of the 2014 Annual Meeting of the American Educational Research Association*, Philadelphia, PA.

Jorion, N., James, K., Schroeder, L. & DiBello, L. V. (2013). *Statistical and diagnostic analyses of student performance on concept inventories*. In *Proceedings of the 2014 Annual Meeting of the American Educational Research Association*, San Francisco, CA.

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Leech, N., & Onwuegbuzie, A. (2007). An array of qualitative data analysis tools: A call for data analysis triangulation. *School Phsychology Quarterly, 22*, 557–584.

Libarkin, J. C., & Anderson, S. W. (2006). The geoscience concept inventory: Application of Rasch analysis to concept inventory development in higher education. In X. Liu & W. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 45–73). Maple Grove, MN: JAM Publishers.

Litzinger, T., Vanmeter, P., Firetto, C., Passmore, L., Masters, C., Turns, S., ... Zappe, S. (2010). A cognitive study of problem solving in statics. *Journal of Engineering Education, 99*, 337–353.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Manktelow, K. (2000). *Reasoning and thinking*. Hove: Psychology Press.

Mestre, J. P., Dufresne, R. J., Gerace, W. J., Hardiman, P. T., & Touger, J. S. (1993). Promoting skilled problem solving behavior among beginning physics students. *Journal of Research in Science Teaching, 30*, 303–317.

Montfort, D., Herman, G. L., Streveler, R. & Brown, S. (2012, October 3–6). Assessing the application of three theories of conceptual change to interdisciplinary data sets. In *Proceedings of the Forty-Second ASEE/IEEE Frontiers in Education Conference* (p. S1B-1–S1B-6), Seattle, WA.

Montfort, D. B., Herman, G. L., Brown, S. A., Matusovich, H. M. & Streveler, R. A. (in press). *Trans-disciplinary patterns in student conceptual understanding of engineering*.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). Columbus, OH: McGraw-Hill.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Pellegrino, J. W., DiBello, L. V., & Brophy, S. P. (2014). 29 the science and design of assessment in engineering education. In A. Johri & B. M. Olds (Eds.), *Cambridge handbook on engineering education research* (pp. 571–600). Cambridge: Cambridge University Press.

Pellegrino, J. W., DiBello, L. V., James, K., Jorion, N., & Schroeder, L. (2011). Concept inventories as aids for instruction: Example applications of a validity framework. In *Proceedings of the Research in Engineering Education Symposium*, Madrid.

Pellegrino, J. W., DiBello, L. V., Miller, R. L., & Streveler, R. A. (2013). Components of a comprehensive approach to validity. In *Proceedings of the 2013 Annual Meeting of the American Educational Research Association*, San Francisco, CA.

Pill, J. (1971). The Delphi method: Substance, context, a critique and an annotated bibliography. *Socio-Economic Planning Sciences, 5*, 57–71.

Porter, L., Garcia, S., Tseng, H. W. & Zingaro, D. (2013). *Evaluating student understand- ing of core concepts in computer architecture*. In *Proceedings of the 18th Annual Conference on Innovation and Technology in Computer Science Education*. Canterbury.

Reiner, M., Slotta, J. D., Chi, M. T. H., & Resnick, L. B. (2000). Naive physics reasoning: A commitment to substance-based conceptions. *Cognition and Instruction, 18*, 1–34.

Rizopoulos, D. (2006). Itm: An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software, 17*, 1–25.

Steif, P. S., & Dantzler, J. A. (2005). A statics concept inventory: Development and psychometric analysis. *Journal of Engineering Education, 33*, 363–371.

Stone, A. (2006). *A psychometric analysis of the statistics concept inventory* (Unpublished doctoral dissertation). University of Oklahoma, Norman, OK.

Streveler, R. A., Miller, R. L., Santiago-Roman, A. I., Nelson, M. A., Geist, M. R., & Olds, B. M. (2011). Rigorous method for concept inventory development: Using the "assessment triangle" to develop and test the thermal and transport science concept inventory (TTCI). *International Journal of Engineering Education, 27*, 968–984.

Taylor, C., Zingaro, D., Porter, L., Webb, K. C., Lee, C. B., & Clancy, M. (in press). *Computer science concept inventories: Past and future*. *Computer Science Education*.

Tew, A. E. (2010). *Assessing fundamental introductory computing concept knowledge in a language independent manner* (Unpublished doctoral dissertation). Georgia Institute of Technology, Atlanta, GA.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*, 535–585.

Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review, 9*, 010116-1.

Webb, K., & Taylor, C. (2014). Developing a pre- and post-course concept inventory to gauge operating systems learning. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, Atlanta, GA.

Wren, D., & Barbera, J. (2014). Psychometric analysis of the thermochemistry concept inventory. *Chemistry & Biodiversity Education Research and Practice, 15*, 380–390.