Journal of Applied Testing Technology, *Vol 17(1), 1-19, 2016*

Assessing Higher-Order Cognitive Constructs by Using an Information-Processing Framework

Philip Dickison[1]*, Xiao Luo[1], Doyoung Kim[1], Ada Woo[1], William Muntean[2] and Betty Bergstrom[2]

[1]National Council of State Boards of Nursing, 111 East Upper Wacker Drive, Suite 2900,
Chicago, IL 60601-4277; pdickison@ncsbn.org
[2]Pearson VUE

Abstract

Designing a theory-based assessment with sound psychometric qualities to measure a higher-order cognitive construct is a highly desired yet challenging task for many practitioners. This paper proposes a framework for designing a theory-based assessment to measure a higher-order cognitive construct. This framework results in a modularized yet unified assessment development system which includes elements spanning from construct conceptualization to model validation. The paper illustrates how to implement this framework by using the construct of nursing clinical judgment. Using this framework, many difficult design decisions can be made with strong theoretical rationales. The framework is also flexible to accommodate modifications and extensions that will be required to be made to the assessment as new knowledge related to the construct is generated over time. The goal of this framework is to provide practitioners with a practical and accessible methodology to assess sophisticated constructs on the ground of cognitive theories of the construct, especially by using technology enhanced items.

Keywords: Assessment Design, Higher-Order Cognitive Construct, Information-Processing Framework, Nursing Clinical Judgment, Technology Enhanced Item

1. Introduction

A body of research has been devoted to the assessment of higher-order cognitive constructs for example, teachers' understandings of rational numbers (Bradshaw, Izsak, Templin, & Jacobson, 2014), high school students' deep conceptual understanding of the Advanced Placement (AP) course content (Huff, Steinberg, & Matts, 2010), English-language proficiency (Chapelle, Grabe, & Berns, 1997; Weir, 2005), complex thinking in mathematics (Graf & Arieli-Attali, in press), argumentation skills (Bertling, Jackson, Oranje, & Owen, 2015), and research and inquiry skills (Sparks & Deane, 2015). These constructs are difficult to measure for two reasons. First, a higher-order cognitive construct is usually an abstract, integrated cognitive practice, such as understanding a concept or creating a verbal or written product. It is hard to attain a unanimous conceptual model that defines underlying mental activities, and also it is likely for theories about the construct to keep evolving in research. The second obstacle is the intrinsic intricacy of higher-order cognitive constructs which often consist of multiple interdependent, sometimes cyclic or chained, cognitive operations. Interactions between one's internal and external knowledge representations may frequently occur, causing one to continuously make intermediate decisions before arriving at a final decision. Such dynamics are difficult to capture with traditional item types, especially the multiple-choice question (MCQ).

To this end, new assessment theories have been introduced, such as evidence-centered design (ECD; Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2003), assessment engineering (AE; Luecht, 2013a), and diagnostic classification modeling (DCM; Rupp, Templin, &

---

*Author for correspondence

Henson, 2012). Both ECD and AE provide a methodology for structured test development with the purpose of accumulating observable evidence to support the claims made about the unobserved trait. DCM, on the other hand, offers a data-analysis approach to obtaining diagnostic information regarding multiple cognitive attributes that constitute a relatively complicated construct. Despite these advancements, designing a theory-based assessment with sound psychometric qualities to measure a higher-order cognitive construct remains a challenging task for many practitioners. This is because all of those theories presume the availability of a sound and comprehensive analysis of the construct which would result in the student model in ECD, the construct map in AE, and the Q-matrix in DCM. Since a robust method of translating cognitive theories into psychometric models is unavailable, construct definition in practice still relies heavily upon subject matter experts' (SMEs) acumen of the construct and their understandings of psychometrics, rather than upon first-hand research results regarding the construct. Likewise, little advice is available pertaining to how to author and score items, especially technology enhanced items (TEIs), on the foundation of cognitive theories and research findings of the construct.

This paper proposes a framework for designing a theory-based assessment to measure higher-order cognitive constructs. This framework results in a modularized yet unified assessment development system (Figure 1). In this framework, a conceptual model is first developed to synthesize cognitive theories and findings about the construct, laying the theoretical grounds for the whole assessment. Subsequently, an assessment model is built upon the conceptual model, which interprets the theories of the construct from the psychometric perspective and translates the theories into psychometric models. The assessment model establishes a psychometric foundation for the assessment that will be used to orchestrate several essential design decisions in the assessment. On the basis of the assessment model, three operational models are built for item authoring (the task model), item scoring (the scoring model), and score interpretation (the mathematical model). Lastly, a validation model is constructed to validate falsifiable design decisions made previously that are related to (a) the correctness of the assessment model and the task models and (b) the appropriateness of the mathematical model and scoring models.

The remainder of this paper will elaborate the function of each model in the framework and illustrate how to implement each model by using the construct of nursing clinical judgment (NCJ). As will be shown below, making the assessment model the center of assessment development enables many difficult design decisions to be made with strong theoretical rationales. Furthermore, the framework is flexible enough to accommodate modifications and extensions that will be required to be made to the assessment as new knowledge related to the construct is generated over time. The goal of this framework is to provide practitioners with a practical and accessible methodology to assess higher-order cognitive constructs, especially by using TEIs.

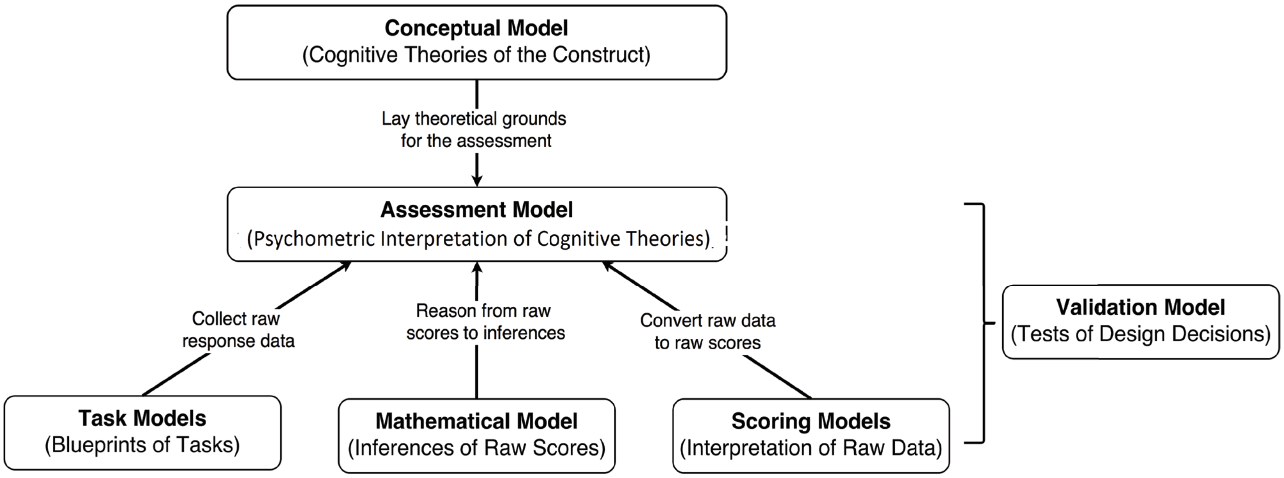


Figure 1. An overview of the information-processing framework for assessing higher-order cognitive constructs.

2. Conceptual Model

Building a conceptual model is the initial step of the framework. The conceptual model synthesizes cognitive theories of the construct, facilitates the understanding of the construct, and hence lays theoretical grounds for the entire assessment. Importantly, this model only defines what the assessment is intended to measure, and other models in the framework address how to measure the construct. In the case of NCJ, this section reviews theories of NCJ that appear in the nursing literature and presents a conceptual model of choice that best suits the purpose of assessing NCJ.

Clinical judgment is recognized by the nursing profession as an essential skill but one that is difficult to measure in an assessment. Phaneuf (2008, p. 1) pointed out that "*nurses must deal with a broad range of issues related to the condition of each patient, including complications and improvements, as well as annotations to clinical records and communications with the physicians.… It is therefore essential for the nurse to have observational and reasoning skills in order to make sound, reliable clinical judgments.*" The pivotal effect of clinical judgment on nursing practice outcomes is well documented in the literature (for a comprehensive review, see Muntean, 2012). Studies have found that a substantial number of the adverse events that hospital inpatients endure may be prevented if decisions had been made using good clinical judgment (Brennan et al., 1991; Hodgetts et al., 2002; Leape, 2000). Contrary to this pressing demand, only one-fifth of employers were satisfied with the decision-making ability of new nurses (Saintsing, Gibson, & Pennington, 2011).

The commonality of different NCJ definitions in the literature suggests that good NCJ requires a set of good observational and intellectual skills (see Tanner, 2006). Nurses with these skills are able to properly identify, associate, and interpret clinical signs, symptoms, and other pertinent data presented in a client situation. Consequently, they are more likely to make good clinical decisions and provide safe and effective care to clients. The literature review (Muntean, 2012) highlighted three NCJ models: (1) the humanistic-intuitive model, (2) the cognitive continuum theory, and (3) the information-processing model. The humanistic-intuitive model defines NCJ as an intuition-like skill acquired through five stages of skill acquisition: the novice, advanced beginner, competence, proficiency, and expert stages (Benner, 2000). While being operationally succinct, this model is oversimplified from a measurement perspective because it fails to illuminate mental operations underlying NCJ sufficiently. Thus, the humanistic-intuitive model is less than fully useful for developing an assessment for NCJ.

The cognitive continuum theory defines NCJ as an adaptive strategy that lies between intuitive and analytic thinking, depending on the context (Harbison, 2001). Specifically, intuitive thinking is invoked for well-structured and familiar decision-tasks while analytic thinking is triggered for ill-structured and unfamiliar decision-tasks. This theory highlights the role of contextual factors in decision-making. Muntean (2012) summarized a wide range of contextual factors related to NCJ that appeared in literature. These include (a) individual factors such as age, education, knowledge, experience, cue recognition, hypothesis updating, communication, emotion, perceptions, confidence, professional orientation, consequence awareness, and personal values, and (b) environmental factors such as task complexity, time pressure, interruptions, area of specialty, and professional autonomy. The broad spectrum of contextual factors reflects the complexity of NCJ and the difficulty in assessing it.

Lastly, the information-processing model defines NCJ as an information processing system, in which multiple cognitive subcomponents of NCJ as well as the order in which they are executed are described. Information-processing models define basic cognitive building blocks of the construct and describe how relevant information is sampled, retrieved and integrated within or across these building blocks (Oppenheimer & Kelso, 2015). This allows models to account for core findings related to the construct as well as for scientific scrutiny to verify the model specification. For this reason, cognitive psychologists who study judgment and decision making are also moving away from expected utility frameworks to cognitive process models (e.g., see Oppenheimer & Kelso, 2015). With respect to NCJ, Muntean (2015) proposed a model with five iterative processes: (1) recognize cues, (2) generate hypotheses, (3) judge hypotheses, (4) take action, and (5) evaluate outcomes (see Figure 2; cf. Elstein, Shulman, & Sprafka, 1978). With this model, it is possible to design tasks to measure each individual process as well as to identify at which process nurses may make errors. Because this model best suits the needs of assessing NCJ, it is adopted as the conceptual model of the NCJ assessment.

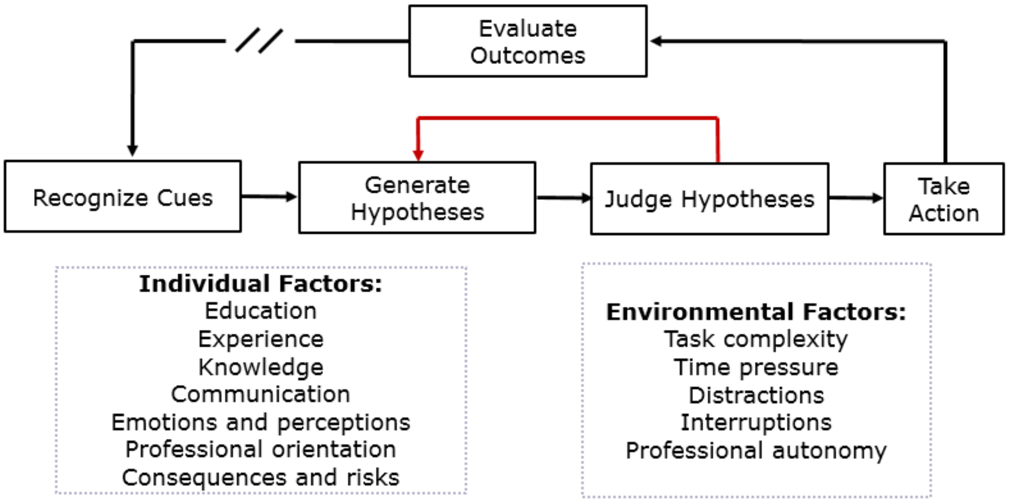


Figure 2. The conceptual model of NCJ from the information-processing perspective (Muntean, 2015)

3. Assessment Model

As mentioned earlier, the conceptual model defines what to measure but not how to measure. Important psychometric issues are not fully addressed in the conceptual model, such as measurement dimensionality, response modeling, item authoring, scoring methodology (especially for TEIs), and so forth. Therefore, an assessment model is developed to interpret the conceptual model from the psychometric perspective, creating a psychometric foundation upon which other models (e.g., task models, scoring models, and the mathematical model) are further developed.

It should be noted that entities in the conceptual model are of two types: cognitive operation and contextual factor. Cognitive operation (e.g., *recognize cues*) is a mental information-processing operation which is directly related to the construct being measured. Cognitive operations are the subjects of measurement about which data will be collected or inferences made. Conversely, contextual factor (e.g., *task complexity*) is a factor that is not an essential constituent of the construct but could influence the outcome of a cognitive operation. Contextual factors are thus the subjects of manipulation in item development in order to control the quality of item production. The assessment model should reflect such distinction.

Also, it is important to distinguish cognitive operation from cognitive attribute. As in the Q-matrix in DCM, a cognitive attribute (e.g., *memorization, deductive reasoning, etc.*) is an individual cognitive skill. Yet, a cognitive process (e.g., *recognize cues*) involves one or multiple cognitive attributes. Both can be the subjects of measurement. When cognitive attributes are measured, scores indicate the mastery level of the cognitive attributes. When cognitive operations are measured, however, scores reflect the ability of completing the cognitive operations to achieve desired outcomes. Because the NCJ is a sophisticated construct which likely involves numerous cognitive attributes, it was decided to measure cognitive operations as specified in the conceptual model, instead of individual cognitive attributes.

As a result, a multilayer assessment model of NCJ is created to establish a psychometric representation of NCJ (Figure 3). Layer 0 (the observation layer) contains two naturally observable entities: *client needs* and *clinical decisions*. The former initiates a NCJ practice, and the latter—whether right or wrong—terminates the practice.

Layers 1–3 (the construct layers) are layers of cognitive analysis of the unobserved construct in the form of a series of cognitive operations. Layer 1 contains a single entity, *clinical judgment,* that encapsulates the entire machinery of NCJ and bridges the two entities in Layer 0. The finer definition of this broad and ambiguous entity is provided in Layer 2 and Layer 3. In Layer 2, the machinery of NCJ is delineated as an iterative process of three cognitive operations: *form hypotheses*, *take actions*, and *evaluate outcomes*. These operations are repeated until the outcome evaluation is perceived as satisfactory by a nurse. In Layer 3, the *form hypotheses* operation is further decomposed into two operations: *recognize cues* and *analyze cues*. The *take actions* operation is also divided into two operations: *prioritize hypotheses* and *generate solutions*. The unobservable entities in Layers 1–3 may generate observable "outcomes" that are measurable and scorable; however, these outcomes differ from the *clinical decision* entity in Layer 0, because they neither directly address the *client needs* nor close the NCJ practice. Furthermore, the *take actions* entity refers to actions taken to address the priority hypothesis instead of client needs; these actions may be identical to the final clinical decisions if the hypothesis is correct, but they are not necessarily identical.

Layer 4 (the context layer) contains a set of contextual factors that may affect the performance of cognitive operations in above layers. There are two types of contextual factors: individual factors (gray circles in Figure 3) vs. environmental factors (white circles in Figure 3). This layer is likely to be frequently modified, as the nature of the profession changes over time or new research findings emerge. Fortunately, entities in this layer are mainly used to assist item development, rather than measure the construct. Modifications in this layer may change the item development practice but leave the score interpretation intact, because Layers 1—3 remain the same.

Compared with the single layer of a chain of cognitive operations in the conceptual model, the assessment

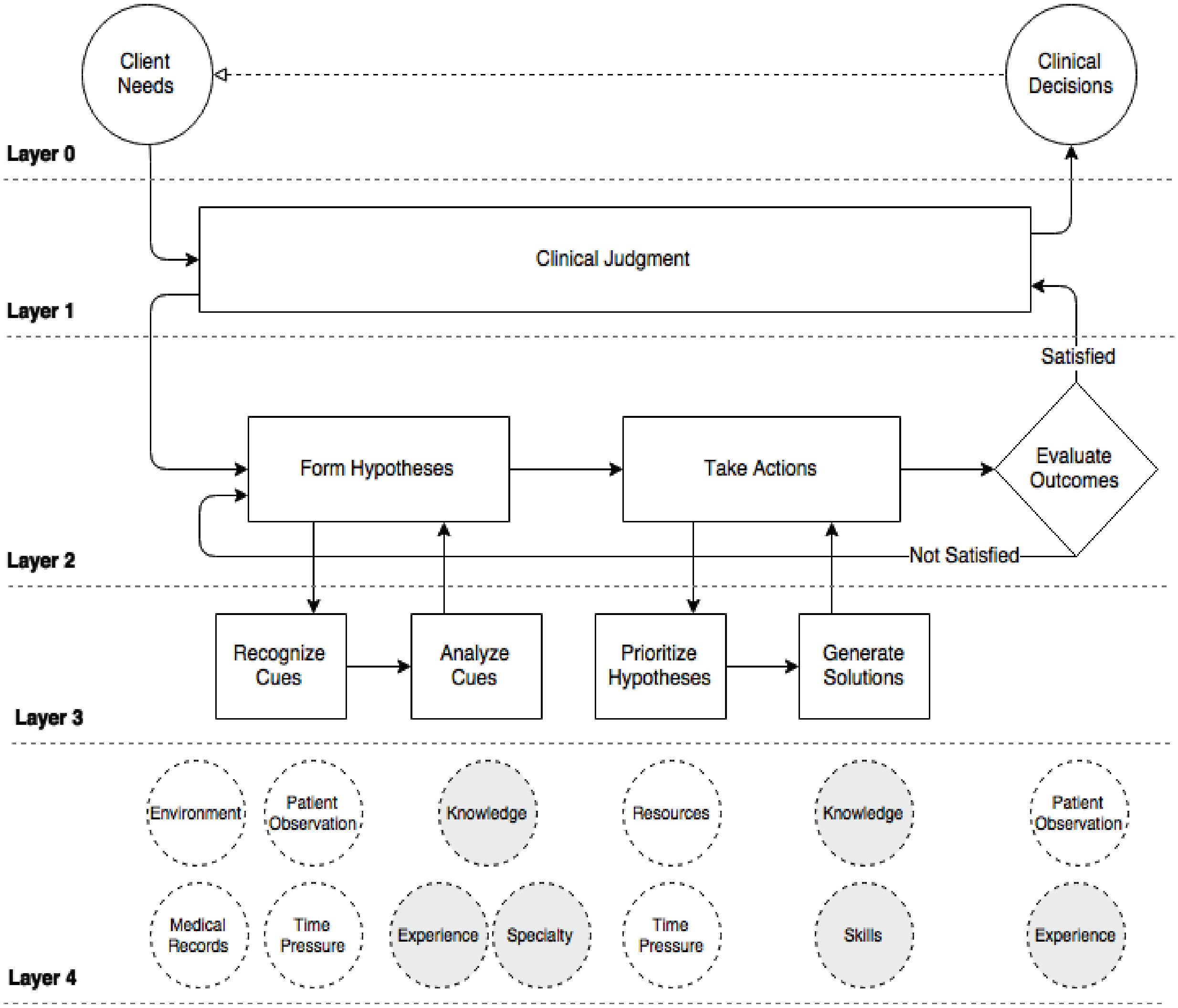


Figure 3. The assessment model of NCJ with the multilayer representation of NCJ.

model represents the construct with various measurement opportunities while still preserving the essential cognitive machinery of NCJ. Given this model, test developers need to make a decision (named the assessment decision in this paper) pertaining to what inferences about NCJ will be made using which layers of the assessment model. For example, it may be decided to model Layer 1 only in order to make an overall pass/fail decision for the licensure purposes, or base the measurement on both Layer 1 and 2 in order to produce multidimensional scores needed for diagnostic purposes. As going deeper through Layers 1–3, greater detail is obtainable regarding the cognition, and finer-grained scores are plausible with advanced item types and measurement models at the expense of increasing operational complexity. Thus, the assessment decision should be made after considering the purpose of the assessment, intended score use, required measurement fineness, resources available, and so forth. How the assessment model is subsequently used as solid theory-based foundation for other essential psychometric decisions will be elaborated in following sections.

4. Mathematical Model

The purposes of a mathematical measurement model are twofold. Firstly, it establishes an avenue of probabilistic reasoning from empirical data to the inferences defined in the assessment decision. Secondly, it defines what types of data are to be collected to realize such reasoning. For instance, when deciding to using only Layer 1 to generate a pass/fail score for the licensure purpose, a unidimensional dichotomous item response theory (IRT) model is likely to be sufficient, with x (empirical data) being the correctness of *clinical decisions* and θ (latent trait) being NCJ proficiency. The relationship between θ and x is modeled as a monotonic probabilistic reasoning by using the Rasch model (Rasch, 1960):

$$P(x=1|\theta,b)=\frac{\exp(\theta-b)}{1+\exp(\theta-b)} \qquad (1)$$

or the 3-parameteter-logistic (3PL) model (Birnbaum, 1968):

$$P(x=1|\theta,a,b,c)=c+\frac{1-c}{1+\exp[-1.7a(\theta-b)]} \qquad (2)$$

where a, b, c are the discrimination, difficulty, and pseudo-guessing parameters of an item, respectively. The monotonic probabilistic relationship between θ and x is displayed in Figure 4 by using three hypothetical items.

The advantage of a unidimensional dichotomous model is its simplicity in terms of conceptualization, implementation, and public communication. Conversely, by being shy of more conceptual detail of NCJ (as shown in Layer 2 and 3), a model of this type may cause biased score interpretation and high item attrition rate. Biased score interpretation is possible because the dimensionality is arbitrarily set using empirical data of a set of items that are presumed to measure the construct coherently. In the extreme case, if all items are authored to measure solely the *form hypotheses* entity or the *take actions* entity rather than a theory-based optimal mixture of both, then the score represents one's *form hypotheses* or *take action* ability. That is not what the theory indicates. As a result, dimensionality analysis is typically conducted for newly developed items to retain items with a similar internal structure of entities in lower layers and eliminate items that do not conform to the arbitrarily-set dimension. Without a rigorous item development methodology that controls the entities in Layer 2 and 3, a large number of newly developed items could be discarded after field test.

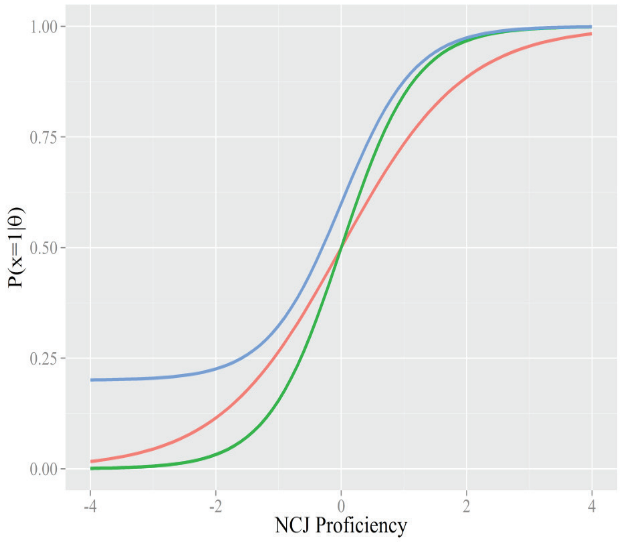


Figure 4. A unidimensional dichotomous model (3PL) of NCJ modeling only Layers 1: a = .5882, b = 0.0, c = 0.0 (red line); a = 1.0, b = 0.0, c = 0.0 (green line); a = 1.0, b = 0.0, c = 0.2 (blue line).

Alternatively, when Layer 2 is added to the assessment decision to produce more information about NCJ, the three-phase information-processing cycle arises and needs to be mathematically modeled. Two types of mathematical models are plausible in this case. First, it is reasonable to posit that the entities in Layer 2 are three consecutive interdependent operations that together lead to a correct clinical decision. This relationship can be expressed by a unidimensional polytomous IRT model, such as the partial credit model (PCM; Masters, 1982):

$$P(x \mid \theta, b, \mathbf{s}) = \frac{\sum_{k=0}^{x} (\theta - b - s_k)}{\sum_{j=0}^{m} \exp \sum_{k=0}^{j} (\theta - b - s_k)}, \quad (3)$$

where $\boldsymbol{s}$ are step parameters for different score categories, and m is the maximum possible score. As is illustrated in Figure 5, the probability of succeeding in all steps increases as the overall NCJ proficiency elevates. Or, viewing entities as separable yet compensatory operations of NCJ, the relationship between empirical data and latent traits can be described by the compensatory multidimensional IRT model (McDonald, 1997):

$$P(x = 1 \mid \boldsymbol{\theta}, \mathbf{a}, d, c) = c + \frac{1 - c}{1 + \exp[-1.7(\boldsymbol{\theta} \cdot \mathbf{a} + d)]} \quad (4)$$

where $\boldsymbol{\theta}$ is a vector of latent abilities residing in three dimensions, $\boldsymbol{a}$ is a vector of discrimination parameters on respective dimensions, and d is the overall difficulty parameter. An example of a two-dimensional model is visualized in Figure 6.

These two models not only differ in mathematical expression, but also, perhaps more importantly, differ in the construct theorization and score reporting. For instance, the polytomous model views clinical judgment as an integrated process with interdependent operations. The nurse has to form correct hypotheses in order to take right actions. Consequently, an overall score is reported to indicate the nurse's ability to perform all cognitive operations in Layer 2 altogether. In contrast, the multidimensional model considers the three cognitive operations as separable dimensions, and it reports a score for each dimension. This approach may be favored in cases where test takers with different skill patterns need to be distinguished. Other advanced models are also plausible as long as it bolsters the assessment decision e.g., a higher-order IRT model, diagnostic classification models, etc. While advanced models are generally more informative, the tradeoffs are increased model complexity, data requirements, and chance of data-model misfit.

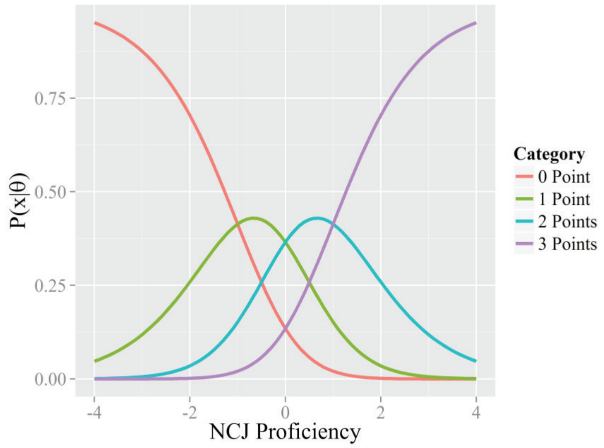


Figure 5. A unidimensional polytomous model (PCM) of NCJ modeling Layers 1–2: b = 0.0, s = [0.0, -1.0, 0.0, 1.0].

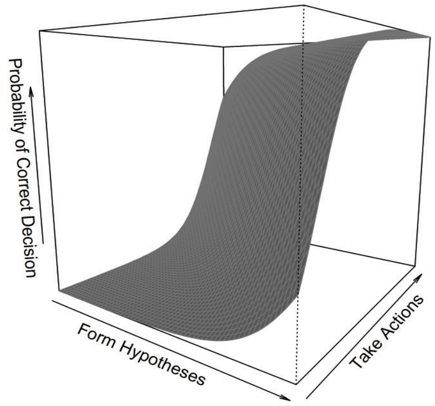


Figure 6. A two-dimensional compensatory model of NCJ modeling Layers 1–2.

5. Task Model

Inferences are made from empirical data, and empirical data are collected from test items. The purpose of the task model is to facilitate the development of highly structured items that elicit responses and generate data from test takers in a consistent manner. Similar to the mathematical model, task models are rooted in the psychometric foundation created by the assessment model. That is, they are created by utilizing entities in the construct layers (Layers 1—3) and the contextual layer (Layer 4) of the assessment model to formulate the abstraction of a family of items that will have similar functionality. Specifically, cognitive operations in the construct layers are used to design tasks and formats, whereas contextual factors in the contextual layer are either manipulated or standardized to achieve desired item characteristics and properties. Actual items are subsequently authored from these "blueprints" of items.

The concept of model-based principled item development has been described in several pioneering studies (Gierl & Lai, 2013; Luecht, Gierl, Tan, & Huff, 2006; Luecht, 2013b; Mislevy, Steinberg, & Almond, 2003). It is embraced in the current framework for three reasons. (1) *Validation*: Explicating the internal cognitive structure of items establishes a rationale from data collection to data utilization. Data are proactively gathered for specific analyses that are set forth in advance, as opposed to being passively analyzed for their given availability. (2) *Item Development*: Authoring items from meticulously crafted/curated task models is more likely to produce items with desired characteristics and properties. (3) *Scoring*: Task models provide useful references for creating scoring protocols so that empirical data are interpreted prescriptively rather than haphazardly.

Diverse operational definitions of task modeling are found in the literature. Gierl and Lai (2013) put a task model in a relatively concrete context with several variables in the question prompt and choice options. Actual values of variables are selected from either a collection of discrete values or a range of continuous values. Mislevy et al. (2003) regarded task modeling as discovering and regulating a set of key variables related to the claims being expectedly made. Luecht (2013b) considered a task model to be a *cognitively oriented specification* of test items, which can be expressed as a series of formulated actions upon objects in a particular context using given tools. Unlike Gierl and Lai's approach, the latter two approaches detach task models from concrete context and content. They grant item developers the autonomy to embody a task model with idiosyncratic context, content and format.

The creation of task models in the current framework is rooted in the assessment model. For instance, a task model can be devised to include cognitive operations in Layer 3 and manipulate factors in Layer 4. Layer 3 entities are used to create tasks and formats, and Layer 4 entities are used to achieve desired item characteristics and properties.

In terms of Layer 3 entities, a task model can be crafted as an integrated or a divided model of those operations. Integrated task models carry multiple cognitive operations in one task, which creates a natural, realistic workflow and interactions/connections between mental activities—for example, designing a task in which the test taker needs to *recognize cues*, *analyze cues*, *prioritize hypotheses*, *generate solutions*, and *evaluate outcomes* altogether. Conversely, divided task models target a single cognitive operation in one task, which requires a less laborious task authoring and scoring procedure but may need a scrupulous data/score synthesis method to combine scores—for example, designing a task to measure whether the test taker *recognizes cues* correctly. Regardless of integrated or divided task models, a balanced mixture of targeted cognitive operations at the assessment level is central to the proper construct representation and unbiased score interpretation (for example, neither "*form hypotheses*" nor "*take actions*" should be over- or under-represented).

Layer 4 outlines critical contextual factors involved in corresponding cognitive operations. Some contextual factors can be controlled to standardize the context, whereas others can be manipulated to achieve targeted task characteristics and properties (e.g., difficulty). Individual factors, on the other hand, cannot be easily manipulated; however, they can be used to clarify the desired interaction between a test taker's internal knowledge representation and the external context. It is suggested that factor conditioning statements be written using structured language, such as:

action [instance] {to [value]} {as [type]}, or (5)
action <a nested factor conditioning statement>. (6)

Action is a verb predefined to a specific type of manipulation. For instance, use *"show"* to mean explicitly giving some information in the prompt, use *"set"* to mean manipulating some information (including holding a variable constant), and use "*imply*" to mean giving some information only when queried or probed for by the test taker. *Imply* can be effectively implemented in TEIs through an interactive interface. *Instance* is a concrete case of a *type* of factor. For example, *room temperature* is an instance of the *environment* cue, and *vital signs* are instances of the *patient observation* cue. The *value* enclosed in the curly

Table 1. Examples of Creating Factor Conditioning Statements by Using Structured Language

Action	Instance	Value (Optional)	Type (Optional)
show	a patient's oral temperature	to 37° C	as patient observation cue
set	location	to emergency room	as environment cue
imply	medical history of diabetes	-	as medical record cue

Table 2. A Hypothetical Task Model in the Pediatric Setting

Cognitive Operation	Factor Conditioning	Expected Behavior
Recognize Cues	Environmental Cues: • Set *location* to *emergency room* • Show *the presence of parent* Patient Observation Cues: • Show *age* to *8-10* • Show *dehydration symptoms* (e.g., dry mucous membranes appear, cool extremities, cap refill 3-4 seconds) • Show/Imply *lethargy* Medical Record Cues: • Show *dehydration symptoms* (e.g., a lower-grade temperature, diarrhea, a poor appetite) • Show/Imply *history of diabetes* • Show/Imply *vital signs* Time Pressure Cue: • Set *time pressure* to *varying with onset of symptoms and current lethargy*	• Recognize *abnormal vital signs* • Recognize *symptoms of dehydration* • Identify *the history of diabetes* • Hypothesize *dehydration* • Hypothesize *diabetes*
Analyze Cues	• Require *knowledge* of *dehydration symptoms* • Require *knowledge* of *diabetes symptoms*	
Prioritize Hypotheses	• Give *vital sign monitors* as *resources* • Set *time pressure* to *vary with vital signs*	• Prioritize *dehydration* • Address *dehydration* • Avoid *glucose*
Generate Solutions	• Require *knowledge* of *dehydration treatment and intervention* • Require *knowledge* of *diabetes treatment and intervention*	
Evaluate Outcomes	Experience: • *Require experience of administering isotonic fluid* Patient Observation Cue: • *Show patient awaking and talking* • *Imply <Set vital signs to varying with action>*	• Check *vital signs* • Check *lethargy*

braces is an optional part used to indicate a specific quantity, range, or nested relationship within another *instance*. For example, the value of a patient's oral temperature can be *37° C*, *36—38° C*, or *vary with age*. More examples of factor conditioning statements are provided in Table 1.

Table 2 presents a hypothetical task model in the pediatric setting which aims to assess whether the test taker is able to make correct clinical decisions using NCJ when exposed to a relatively realistic and common client situation in the emergency room. The factor conditioning column uses the structured language to regulate contextual factors associated with corresponding cognitive operations, which facilitates the creation of the content and context of items. The expected behavior column indicates the expected behavior for the cognitive operations, which facilitates the creation of task format and scoring protocol. Together, item production is expectedly well controlled by the information in the task model. Appendix I and II present a 3-item item-set authored from this task model.

This task model is very useful when authoring TEIs because it regulates the structure but not the implementation of items. For example, it is ascertained from Table 2 that the client shows dehydration symptoms. This can be implemented via textual descriptions in MCQ format (see Appendix I) or via multimedia content (e.g., images or videos) in TEIs (see Appendix II). Compared with text, multimedia is more realistic and accurate in terms of embedding cues in items. Another example is "imply history of diabetes," where the history of diabetes is shown only when the test taker proactively looks for the client's medical history. This is difficult to implement in conventional item types, because all information is static with limited interactivity in items. However, since TEIs allow far more interactivity, this can be implemented by creating an icon in the testing interface that is linked to show the medical history of the client. Compared with showing the medical history directly, this implementation assesses whether the test taker is aware of the importance of medical history in the scenario, which is part of NCJ construct. The principle here is to create proper implementation to fit the content requirement, as opposed to writing proper content to fit the implementation. The same principle is applicable to response collection in TEIs. Depending on the nature of expected behaviors, responses can be collected by a variety of ways in addition to multiple-choice, such as ordering, drag-and-drop, and hot-spot.

To summarize, the task model controls crucial factors in item production, yet still leaves plenty of room for developers to maneuver the content, context, format, and presentation of items. Regulations imposed upon item production, as indicated in the task model, are derived from the assessment model so that item content is linked to the measurement purpose. This ensures that items are developed to tap key measurement components in well-controlled contexts. TEIs are created when the conventional implementation does not suffice to condition contextual factors or collect expected responses.

7. Scoring Model

After collecting raw item responses, responses need to be converted to data that the mathematical model expects. The purpose of the scoring model is to process the raw responses and result in meaningful numeric values for variables in the mathematical model. Each task model should be coupled with a scoring model. The simplest example is the following scoring model for MCQs:

$$x = \begin{cases} 1, \text{ when response } \textit{matches} \text{ key} \\ 0, \text{ otherwise} \end{cases} \tag{7}$$

Another example is essay scoring rubrics which rely on human raters to evaluate the raw response against a set of rules and standards. As the item format becomes more complex and more open-ended, the scoring model becomes increasingly complex as well. It is critical to control for biases and unfairness in scoring under these circumstances. As a result, scoring models should be created from a sound methodology to establish a valid connection between observed data and interpreted (scored) data.

Luo, Becker, Sutherland, and de Jong (2015) proposed an evidence-based scoring (EBS) framework for scoring a variety of item types with following steps:

1. Develop *a construct model* that outlines key *scoring variables* used to bridge empirical data and inferences to be made;
2. Parse the observed responses into *evidentiary objects* (see Table 3);
3. Analyze the structural and dimensional relationship among *evidentiary objects*;
4. Assign raw scores to the *evidentiary objects* in terms of *scoring variables*;
5. Synthesize raw scores of *evidentiary objects* by the structural relationship to intended construct dimensionality.

The EBS framework can be used to develop scoring models for the current framework. For example, suppose it has been decided to use the PCM (modeling Layer 1 and Layer 2 in the assessment model) to produce a unidimensional score for the NCJ assessment. The scoring model for the sample item (Appendix I) would feature a hierarchical architecture consisting of a master variable (*clinical judgment*) and three scoring variables beneath it (*form hypotheses*, *take actions*, and *evaluation outcomes*). Evidence regarding scoring variables would be gathered from observed item responses and used to make

Table 3. Evaluation of Evidentiary Objects in the Sample Item

Evidentiary Object	Scoring Variable			Raw Score
	Form Hypotheses	Take Actions	Evaluate Outcomes	
MCQ #1	1	0	0	See Equation 7
MCQ #2	0	1	0	See Equation 7
MCQ #3	0	0	1	See Equation 7

inferences about the master variable through the PCM. Because the sample item (see Appendix I) is authored from a divided task model (i.e., one MCQ targeting only one cognitive operation), it yields a map of evidentiary objects and scoring variables as in Table 3.

In this example, a binary scale is used to indicate whether an evidentiary object contributes to a scoring variable. For instance, a correct response to MCQ #1 accumulates one bit of evidence to the scoring variable of *form hypotheses* but not to the other two scoring variables. The raw score is obtained by comparing the response to the key (see Equation 7).

When synthesizing raw scores of these three MCQs, one needs to be aware of the precedence assumption made in the mathematical model. That is, a point is awarded for a scoring variable only when all preceding steps are correct. For instance, an incorrect response to MCQ #1 would void points in MCQ #2 and MCQ #3. A raw score vector (0, 1, 1) corresponds to a final score of 0 instead of 2. Mathematically, this scoring rule is given by:

$$X = \sum_{i=1}^{3} \prod_{j=1}^{i} x_j = x_1 + x_1 x_2 + x_1 x_2 x_3 \quad (8)$$

where x_j is the score on j-th scoring variable, and X is the final score on this item. Whether or not this precedence assumption is rational can be validated by using the validation model described in the next section. If this assumption is found to be invalid, it would be relaxed, followed by modifications of the mathematical model and the scoring model.

Although the sample item (see Appendix I) includes only MCQs, it is reasonable to employ other item types as long as they properly elicit intended responses from test takers (see Appendix II). As the expected response shifts from closed-ended to open-ended, identification of evidentiary objects takes more effort and care. An extreme example would be constructed-response (CR) items. Evidentiary objects for a CR item could be keywords, phrases, and sentences that are carefully identified using techniques like natural language processing and textual data mining. In some other occasions, a 3-point scale may be used to indicate whether an evidentiary object contributes positive, neutral, or negative evidence to a scoring variable. If the task is authored from an integrated task model, the scoring rule could be much more complex than that in Table 3 and Equation 8, since one evidentiary object may contribute to multiple scoring variables.

8. Validation Model

Assessment validation is the cornerstone of test development and refers to a comprehensive ongoing investigation of the correctness of score interpretations and consequences of score uses (Kane, 2006; Messick, 1995). As a result, it is essential to develop a validation model to test various falsifiable design decisions for the assessment. In particular, the validation model described below primarily addresses validating the assessment model, the mathematical model, task models, and scoring models. While positive validation results would provide valuable validity evidence, they would not constitute a comprehensive validation of the assessment.

Although use of a mixture of qualitative and quantitative validation methodologies is strongly encouraged, we primarily focus on building a data-driven validation model for the example described above for purposes of demonstration. The assessment model and the mathematical model can be falsified in various ways. First, a model-data misfit would signal the incorrectness of the assessment model or the mathematical model. Model fit should be judiciously examined using a method of choice (Ames & Penfield, 2015). Second, it is advisable to inspect the dimensionality of the assessment and correlational relations of cognitive entities to test whether the modeled cognitive entities exhibit a reasonable degree of association or distinctiveness. Possible analytic methods are multidimensional IRT modeling, structural equation modeling, factor analysis, and path analysis. A hypothetical diagram like Figure 7 would be expected from these analyses.

Item analysis is central to validate task models and scoring models. The statistics of individual items (e.g., item difficulty, item fit, point-biserial correlation, etc.) can be inspected to detect misbehaving items. Meanwhile, items that are derived from the same task model (i.e., sibling items) can be analyzed collectively. A high level of internal consistency among sibling items is expected; otherwise, either some items are not correctly authored from the task model, or the task model demands stricter regulations over contextual factors. If a task model yields a large number of misbehaving items, then the task model and the corresponding scoring model should be revised.

9. Conclusion

With fast-developing technologies and test stakeholders' increasing demands for advanced assessments, psychometricians are charged with the responsibility for innovating assessment theories to tap sophisticated higher-order cognitive constructs. Indeed, this paper is focused on measuring NCJ, a construct that is still in the process of investigation and definition. However, current knowledge about NCJ derived from its own specialized field of study can be used to develop and refine assessment of the construct. Indeed, lack of innovation in assessment theory can hamper the measurement of higher-order cognitive constructs, largely because of the challenges associated with effectively linking advancements in other fields (such as cognitive psychology, learning theory, and others) with assessment. As a result, the framework for assessing higher-order cognitive constructs introduced in this paper highlights how cognitive theories can be effectively translated into psychometric models. Further, since theories are likely to evolve over time, the framework is also designed to be flexible to accommodate modifications and extensions as new findings about the construct emerge in research.

Compared with conventional assessment design methodologies, this framework has three advantages. The first is the centralization of the assessment design. As illustrated, this framework results in a relatively comprehensive, self-contained assessment system consisting of models regarding construct, measurement, item development, scoring, and validation. The assessment model is at

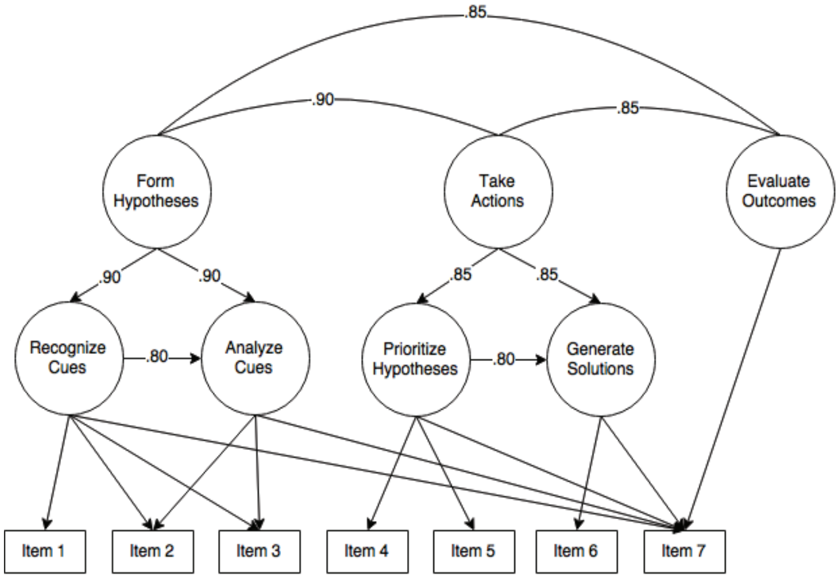


Figure 7. A hypothetical diagram of relationships between cognitive entities.

the heart of the system, establishing a psychometric foundation based upon cognitive theories of the construct and orchestrating essential psychometric design decisions in the assessment. This ensures the entire assessment is constructed with a unified purpose. Similar ideas of the unified assessment design pattern can be found in ECD (Mislevy et al., 2003) and AE (Luecht, 2013a).

The second advantage is modularization of the assessment design. While being integrated to achieve the same purpose, each model has an exclusive function and responsibility within the entire framework. This modularized design pattern allows flexible modifications. For example, a new mathematical model can be applied without interfering with task models or scoring models (i.e., item authoring and scoring rules remain the same). Likewise, as new knowledge about the construct is obtained, task models and scoring models can be revised without modifying the mathematical model (i.e., score interpretation remains the same).

The third advantage is the embrace of an information-processing perspective. As illustrated in the assessment model, higher-order cognitive operations (e.g., *clinical judgment*) typically involve multiple sub-operations and invoke numerous cognitive attributes. If the construct is rudimentarily theorized as a weighted sum of cognitive attributes, it is likely to result in unmanageable dimensionality and overcomplicated task development, due to the quantity and intricacy of those attributes. Conversely, the information-processing perspective focuses the assessment on cognitive operations, regardless of the cognitive attributes invoked in each operation. Scores are consequently used to extrapolate a test taker's probability of completing the operation to achieve desired outcomes. Cognitive attributes, on the other hand, are mainly used to facilitate item development and scoring. In this way, the resultant assessment retains a parsimonious structure.

Additionally, the paper demonstrates how this framework can be used to develop TEIs. Essentially, task models are created to regulate contextual factors in each cognitive operation included in the task but leave the implementation of factor conditioning and response collection open-ended. This highlights how TEIs are needed only when conventional item types are unable to measure the construct of interest. Following this approach, TEIs can be developed to fit specific content requirements, as opposed to content being written to fit the item format requirements. To this point, a frequent criticism of many TEIs is that they may not necessarily offer significant measurement advantages beyond test taker engagement. As described above, the item development method in this framework can help to ensure that TEIs are developed meaningfully to tap sophisticated construct components rather than to simply add visual appeal, minimizing the risk of inadvertently introducing construct-irrelevant variances in scores.

One limitation in this paper is that we assume the conceptual model of choice represents NCJ sufficiently. That is, the information-processing model is considered to subsume the other two models, if intuition can be considered an expedited information-process practice where one is so familiar with the problem that swiftly completes the whole process. However, as one reviewer notes, no evidence is provided to justify this assumption. While discussion of this is beyond the scope of this paper, our focus as measurement professionals is on building assessments on the basis of cognitive theories rather than investigating cognitive theories themselves. Additionally, justification of this assumption can be included in the validation model. If the assessment is built upon false assumptions, the analysis in the validation model should reveal that error.

Put together, this framework is expected to give practitioners the capacity to design an assessment to measure higher-order cognitive constructs. Also, it can be employed to analyze an existing assessment in order to (1) obtain a better understanding of design decisions that were previously made implicitly, (2) improve task development by constructing or refining task models, and (3) collect assessment validation evidence by implementing a validation model.

10. Acknowledgements

The authors would like to acknowledge Nicole Williams, MSN, RN, and Lauren Derby, MSN, RN, for developing the hypothetical task model and the sample items. We would also like to thank Marie Lindsay, MS, RN, and Karen Sutherland, PhD, RN, for their contribution to an earlier version of the clinical judgment model. This research would not have been possible without the guidance and clinical expertise from these nurses.

11. References

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, *34*, 1 –10.

Benner, P. D. (2000). *From novice to expert: Excellence and power in clinical nursing practice, commemorative edition*. Upper Saddle River, NJ: Prentice-Hall.

Bertling, M., Jackson, G. T., Oranje, A., & Owen, V. E. (2015). Measuring argumentation skills with game-based assessments: Evidence for incremental validity and learning. In *Artificial Intelligence in Education* (pp. 545–549). New York, NY: Springer.

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, *33*, 2-14.

Brennan, T. A., Leape, L. L., Laird, N. M., Hebert, L., Locaio, A. R., Lawthers, A. G., Hiatt, H. H. (1991). Incidence of adverse events and negligence in hospitalized patients: Results of the Harvard Medical Practice Study I. *New England Journal of Medicine*, *324*, 370–376.

Chapelle, C., Grabe, W., & Berns, M. S. (1997). *Communicative language proficiency: Definition and implications for TOEFL 2000*. Princeton, NJ: Educational Testing Service.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Graf, E. A., & Arieli-Attali, M. (in press). Designing and developing assessments of complex thinking in mathematics for the middle grades. *Theory Into Practice*.

Gierl, M. J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, *32*, 36–50.

Harbison, J. (2001). Clinical decision making in nursing: Theoretical perspectives and their relevance to practice. *Journal of Advanced Nursing*, *35*, 126–133.

Hodgetts, T. J., Kenward, G., Vlackonikolis, I., Payne, S., Castle, N., Crouch, R., … Shaikh, L. (2002). Incidence, location and reasons for avoidable in-hospital cardiac arrest in a district general hospital. *Resuscitation*, *54*, 115–123.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, *23*, 310–324.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (Vol. 4, pp. 17-64). Lanham, MD: Rowman & Littlefield Publishers.

Leape, L. L. (2000). Institute of Medicine medical error figures are not exaggerated. *Journal of American Medical Association*, *284*, 95–97.

Luecht, R. M. (2013a). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology, 14,* 1-38.

Luecht, R.M. (2013b). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 59–76). New York: Routledge.

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006). *Scalability and the development of useful diagnostic scales.* Paper presented in the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Luo, X., Becker, K., Sutherland, K., & de Jong, J. (2015). *Constructing a framework for scoring innovative test items.* Paper presented in the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257–269). New York, NY: Springer.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*, 6-20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–62.

Muntean, W. J. (2012). *Nursing clinical decision-making: A literature review*. Paper commissioned by the National Council of State Boards of Nursing.

Muntean, W. J. (2015). *Evaluating clinical judgment in licensure tests: Applications of decision theory.* Paper presented in the annual meeting of the American Educational Research Association in Chicago, IL.

Oppenheimer, D. M., & Kelso, E. (2015). Information processing as a paradigm for decision making. *Annual Review of Psychology*, *66*, 277-294.

Phaneuf, M. (2008). *Clinical judgment—An essential tool in the nursing profession*. Retrieved on June 8, 2015 from http://www.infiressources.ca/fer/Depotdocument_anglais/Clinical_Judgement%E2%80%93An_Essential_Tool_in_the_Nursing_Profession.pdf

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen, Denmark: Danish Institute for Educational Research*, 1960.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications*. New York: Guilford.

Saintsing, D., Gibson, L. M., & Pennington, A. W. (2011). The novice nurse and clinical decision-making: How to avoid errors. *Journal of Nursing Management*, *19*, 354-359.

Tanner, C. (2006).Thinking like a nurse: A research based model of clinical judgment in nursing. *Journal of Nursing Education*, *45,* 204–212.

Sparks, J. R., & Deane, P. (2015). Cognitively based assessment of research and inquiry skills: Defining a key practice in the English language arts. *ETS Research Report Series*.

Weir, C. J. (2005). *Language testing and validation*. London, UK: Palgrave Macmillan.

Appendix I

An 8-year-old client with a history of diabetes presents to the emergency room with his mother, who reports that the child has not been feeling well for the last two days. She states he has a low-grade temperature, diarrhea, and a poor appetite. Today, the child reports he is feeling dizzy and that his head hurts. The mother also reports that he is refusing to eat or drink anything. Client vital signs upon arrival are pulse–162 beats/minute, respirations–26 breaths/minute, blood pressure–78/42 mmHg, temperature-100.3° F orally and blood serum glucose-75mg/dL. The client is admitted to the hospital, and an intravenous line is placed with 0.9% normal saline infusing at 50mL/hr. The nurse notes that the child is responsive to questions but appears lethargic. The mucous membranes appear dry, extremities are cool, and capillary refill is 3-4 seconds.

1. Which of the following orders can the nurse anticipate?
   a. Administer an intravenous fluid bolus of isotonic fluid (Key).
   b. Offer a cola beverage.
   c. Administer acetaminophen.
   d. Administer oxygen via nasal cannula.

The nurse re-evaluates the client after two hours from the initial admission. The child is awake and talking, extremities remain cool, and capillary refill is 2-3 seconds. The client is asking to drink something. Client vital signs are pulse–152 beats/minute, respirations–22 breaths/minute, blood pressure–82/46 mmHg, temperature-100.2° F orally. Laboratory values: electrolytes, within normal limits; blood serum glucose, 80mg/dL.

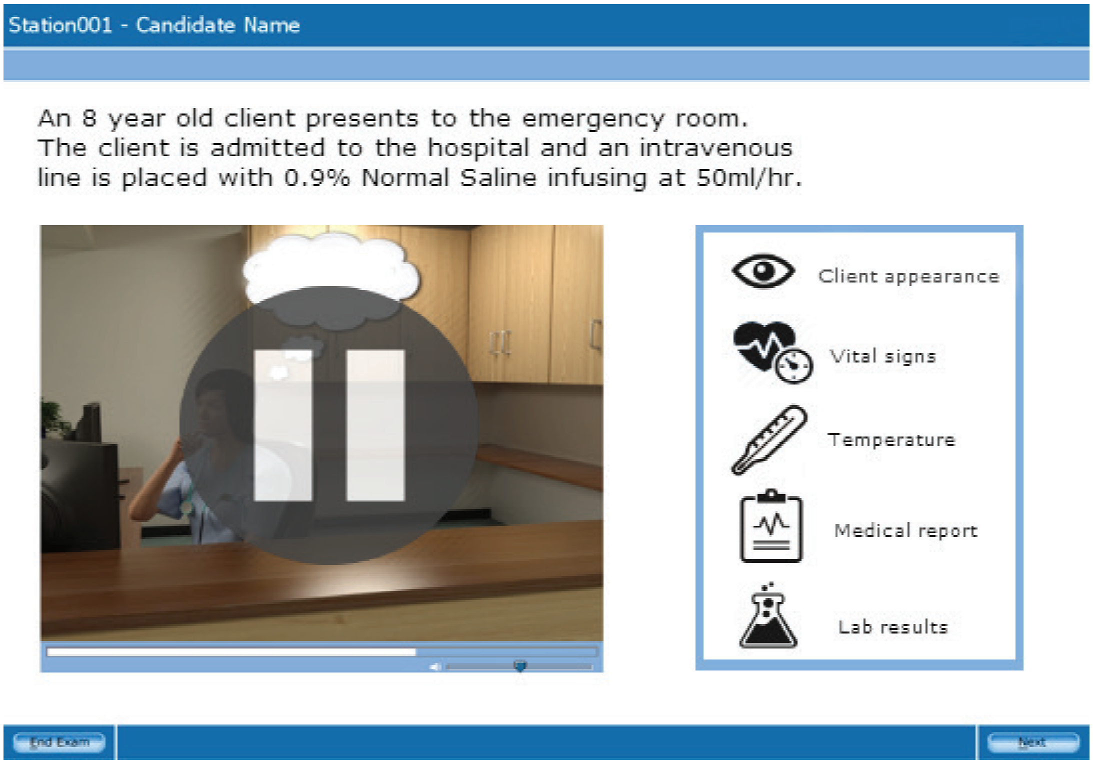
2. Which of the following actions should the nurse take?
   a. Administer an intravenous fluid bolus of isotonic fluid (Key).
   b. Administer insulin.
   c. Increase the 0.9% normal saline intravenous fluid rate.
   d. Discontinue the intravenous line.

3. The nurse re-evaluates the client after four hours from the initial admission. Which of the following findings indicate that the client's treatment has been effective?
   a. blood glucose of 85mg/dL
   b. pulse of 100 beats/minute (Key)
   c. respiration rate of 20 breaths/minute
   d. oral temperature of 100° F

Appendix II

In this example, a hypothetical technology enhanced item is presented. This item is created using the same task model described in the paper. Test takers will read the text on the screen, watch the video/audio clip (if available), and click icons in the blue box to obtain corresponding information (if available). Test takers will use their ability to retrieve, interpret and synthesize information presented in different formats (e.g., text, images, sound, etc.).

This item consists of three phases: the initial admission of the client to the emergency room (Phase 1), two hours after the initial admission (Phase 2), and four hours after the initial admission (Phase 3). Each phase is coupled with one question to answer. Test takers will only see the following five images of testing interfaces. Three tables shown below are to explain what is contained in the video/audio clip and clickable icons. They will not be presented to the test takers.

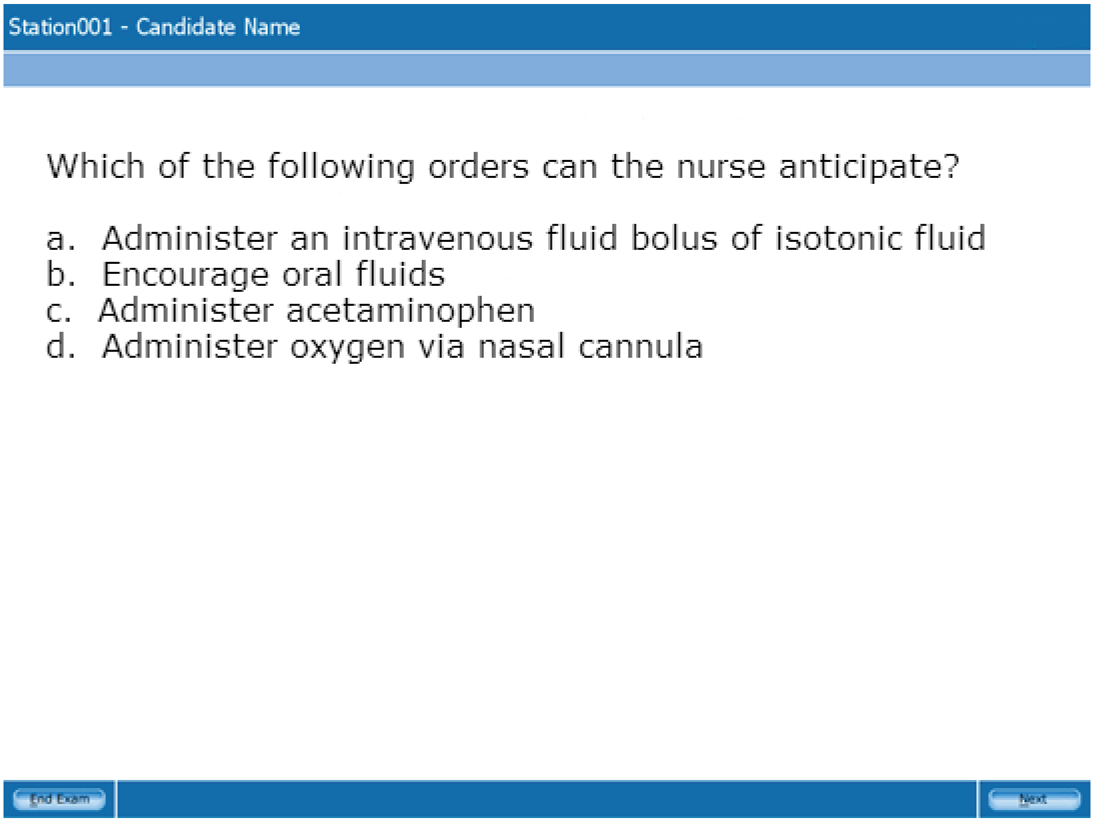


An 8 year old client presents to the emergency room.
The client is admitted to the hospital and an intravenous
line is placed with 0.9% Normal Saline infusing at 50ml/hr.

Figure 1. Testing interface for Phase 1 (the initial admission of the client to the emergency room).

Table 1. Information in the Testing Interface for Phase 1

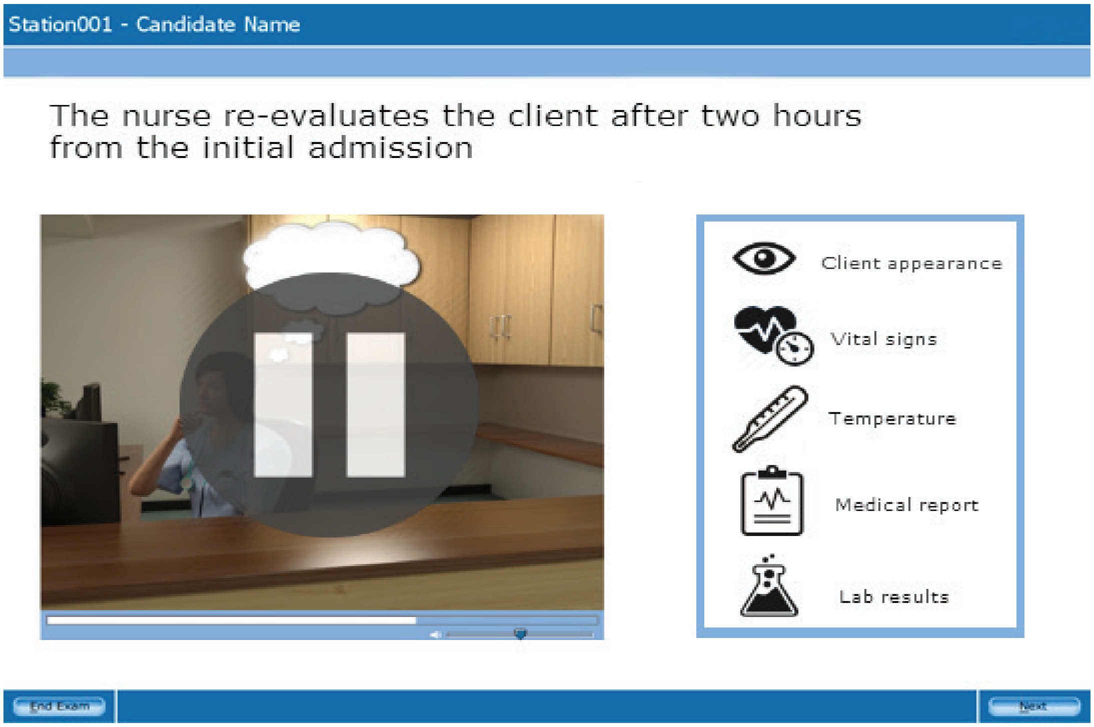
Interface Icon	Content
Video/audio clip	A conversation as follows: • <u>*Nurse:*</u> *Hi, what brings you to the emergency room today?* • <u>*Mother:*</u> *My child has not been feeling well for the last two days. He has a low-grade temperature, diarrhea, and a poor appetite. He is also refusing to eat or drink anything.* • <u>*Child:*</u> *Mommy, I feel dizzy and my head still hurts.* The child is responsive to questions but appears lethargic The mucous membranes appear dry, extremities are cool Capillary refill is 3-4 seconds
Client Appearance (image)	The child appears lethargic The mucous membranes appear dry
Vital signs (image)	Pulse-162 beats/minute Respirations-26 breaths/minute Blood pressure-78/42 mmHg
Temperature	100.3F orally
Medical Report	A history of diabetes
Lab results	No results are available


Station001 - Candidate Name

Which of the following orders can the nurse anticipate?

a. Administer an intravenous fluid bolus of isotonic fluid
b. Encourage oral fluids
c. Administer acetaminophen
d. Administer oxygen via nasal cannula

Figure 2. Testing interface for Question #1.

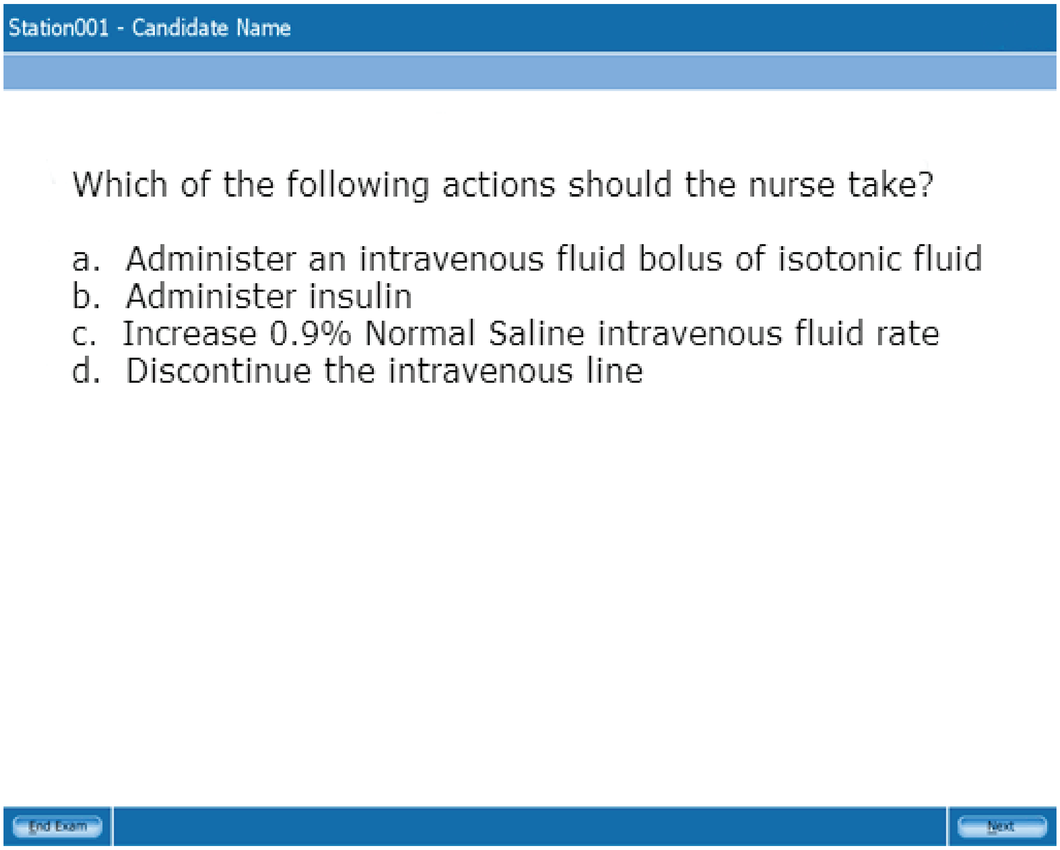

Station001 - Candidate Name

The nurse re-evaluates the client after two hours from the initial admission

Figure 3. Testing interface for Phase 2 (two hours after the initial admission).

Table 2. Information in the Testing Interface for Phase 2

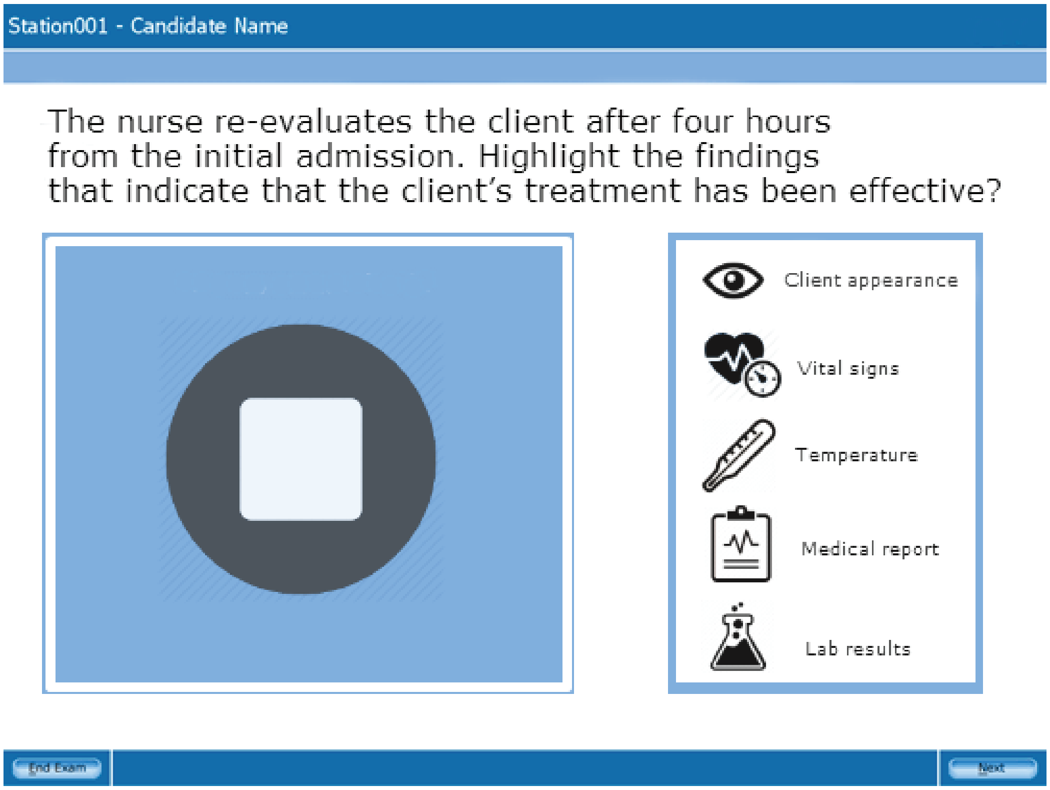
Interface Icon	Content
Video/audio clip	The child is awake and talking, extremities remain cool The client is asking to drink something Capillary refill is 2-3 seconds
Client appearance (image)	The child is awake and talking
Vital signs (image)	Pulse-152 beats/minute Respirations-22 breaths/minute Blood Pressure-82/46 mmHg
Temperature	100.2F orally
Medical report	A history of diabetes
Lab results	Electrolytes within normal limits Blood serum glucose 80mg/dL



Which of the following actions should the nurse take?

a. Administer an intravenous fluid bolus of isotonic fluid
b. Administer insulin
c. Increase 0.9% Normal Saline intravenous fluid rate
d. Discontinue the intravenous line

Figure 4. Testing interface for Question #2.


Station001 - Candidate Name

The nurse re-evaluates the client after four hours from the initial admission. Highlight the findings that indicate that the client's treatment has been effective?

Figure 5. Testing interface for Phase 3 (four hours after the initial admission) and Question #3.

Table 3. Information in the Testing Interface for Phase 3

Interface Icon	Content
Video/audio clip	No video clips are available
Client appearance	Not available
Vital signs (image)	Pulse-100 beats/minute Respirations-20 breaths/minute Blood Pressure-92/64 mmHg
Temperature	100F orally
Medical report	A history of diabetes
Lab results	Electrolytes within normal limits Blood serum glucose 85mg/dL