

# Original Investigations

# JAMIA

## Model Formation ■

## Extending the LOINC Conceptual Schema to Support Standardized Assessment Instruments

---

THOMAS M. WHITE, MD, MS, MA, MICHAEL J. HAUAN, MD, MPH, MTS, MA

**Abstract Objective.** To extend the Clinical LOINC (Logical Observation Identifiers, Names, and Codes) semantic schema to support (1) the representation of common types of assessment instruments and (2) the disambiguation of versions and variants that may have differing reliability and validity.

**Design.** Psychometric theory and survey research framework, plus an existing tool for implementing many types of assessment instruments (Dialogix), were used to identify and model the attributes of instruments that affect reliability and validity. Four modifications to the LOINC semantic schema were proposed as a means for completely identifying, disambiguating, and operationalizing a broad range of assessment instruments.

**Measurements.** Assess the feasibility of modeling these attributes within LOINC, with and without the proposed extensions.

**Results.** The existing LOINC schema for supporting assessment instruments was unable to consistently meet either objective. In contrast, the proposed extensions were able to meet both objectives, because they are derived from the Dialogix schema, which already performs those tasks.

**Conclusion.** These extensions to LOINC can facilitate the use, analysis, and improvement of assessment instruments and thereby may improve the detection and management of errors.

■ *J Am Med Inform Assoc.* 2002;9:586–599. DOI 10.1197/jamia.M1033.

---

Affiliations of the authors: New York State Office of Mental Health, New York, New York (TMW); University of Missouri School of Medicine, Columbia, Missouri (MJH).

This work was supported in part by the Research Foundation for Mental Hygiene via a Fellowship in Psychiatric Informatics at the New York State Psychiatric Institute; NIMH Grant MH54601; and the New York State Office of Mental Health.

Correspondence and reprints: Thomas M. White, MD, MS, MA, Assistant Director, Bureau of Evidence Based Medicine and Practice Guidelines, New York State Office of Mental Health, 330 Fifth Ave, 9th Floor, New York, NY 10001; e-mail: <tw176@columbia.edu>.

Received for publication: 10/13/01; accepted for publication: 6/24/02.

Clinicians and researchers strive to optimally diagnose patients and to track the changes in their health and how such changes impact their lives. Structured assessment instruments can be used to reliably measure a broad range of attributes of patient health and status. Unfortunately, the lack of standard terminologies or coding systems for these instruments limits sharing, re-use, and analysis of changes in these measures. Moreover, some instruments are locally modified in ways that may change their meaning and accuracy, such that it might be inappropriate to compare data collected from different versions of the instruments.

The use of concept-oriented terminologies can significantly improve the sharing and re-use of the clinical information they encode.<sup>4,5</sup> Computer-mediated decision support, outcomes assessment, and quality assurance systems require that clinical data be encoded using such terminologies. LOINC (Logical Observation Identifiers, Names, and Codes) is one such coding system.

LOINC has been extended to represent and store data from standardized nursing instruments.<sup>6</sup> However, other types of instruments, such as structured interviews and diagnostic algorithms, pose additional challenges that require extensions to LOINC. Moreover, LOINC needs to be extended to address the fact that different versions or non-standard implementations of questions within standardized instruments can change their meaning, even though many clinicians and researchers mistakenly assume that they are equivalent.

The common assumption that all variant uses of an instrument are identical can lead to medical errors and limit their detection. For example, imagine that a patient's symptoms are repeatedly measured with versions of an instrument whose reference ranges are not the same. Reference ranges for different versions can have different widths, and their baselines can shift. For example, the Brief Psychiatric Rating Scale (BPRS)<sup>1</sup> measures symptoms on a fuzzy scale from "not present" to "extremely severe." Some variants do not clarify what constitutes "mild" or "severe" in the context of each symptom. Thus, young clinicians who have not seen the full range of disease might think that a patient with mild disease actually has extreme disease. In such cases, patients might appear to be getting worse (or better), when they are actually remaining the same (a type I error). Similarly, clinicians who routinely work with certain patient populations might consider symptoms as mild when the standards indicate that they are severe. This can result in type II errors in which real changes are acci-

dentally missed. Unfortunately, without the ability to distinguish among instruments and estimate their real-world reliability and reference ranges, these types of medical errors will go undetected.

Efforts to extend LOINC to support standard assessment instruments can benefit from psychometric and survey theory. Many volumes have been written on the theory and best practices of survey research,<sup>7-10</sup> writing questions,<sup>11,12</sup> designing scales and instruments,<sup>10,13</sup> and psychometrically evaluating them.<sup>14</sup> Unfortunately, the task is not easy, as indicated by the many publications bemoaning the poor quality of health survey instruments.<sup>15-19</sup> Survey methodology indicates<sup>9</sup> that variables must be defined along three axes: (1) the conceptual definition, which indicates what is being measured; (2) the operational definition, which is the actual question asked; and (3) the variable definition, which includes the answer options, associated internal coding values, and validation criteria.

Pragmatically, the concept definition is the researcher's hypothesis of what construct or concept is being measured by the operational and variable definitions. Until psychometric measurement studies are performed, the concept definition is only a best guess. Survey methodology research has shown that even minor changes in the operational and variable definitions can dramatically alter which construct is really assessed, and the reference range and accuracy with which it is measured.<sup>10,13,14,20</sup> Since one goal of LOINC is to assign a unique code for each measurable entity with a significantly different clinical meaning or reference range, LOINC needs to address and model the operational and variable definitions of items within assessment instruments. The current LOINC extensions for assessment instruments do model the conceptual definition and naming of variables but are not robust in modeling the operational and variable definitions.

This article reviews the current extensions added to LOINC to support assessment instruments and highlights their limitations. Then it presents an overview of the Dialogix system, which supports a broad range of assessment instruments. Next the article addresses the modeling needs of survey questions by discussing the types of modifications that can result in significant changes in meaning or reference range. Then the article shows how the Dialogix model can be adapted to meet LOINC's needs, and validates the feasibility of such an approach. Finally, the benefits and impact of such changes are discussed.

## Background

### LOINC

LOINC, the terminology for Logical Observation Identifier Names and Codes, was initially developed to facilitate the identification and storage of laboratory tests,<sup>21</sup> with a goal of being a general scheme for naming all possible patient-related observations.<sup>22</sup> The Clinical LOINC semantic model has been extended to classify standardized nursing assessment instruments.<sup>6</sup>

The objective of LOINC is to create universal, unambiguous, observation (measurable entity) identifiers for use in data exchange standards, such as HL7,<sup>23</sup> that send information as name-value pairs.<sup>22</sup> Each measurable entity is classified across six axes, corresponding to six main fields within the LOINC table: *Component*, *Property*, *Timing*, *System*, *Scale*, and *Method*. The LOINC database can store additional information about each test in supplemental tables.

Fully specified LOINC names are supposed to contain sufficient information to disambiguate similar but distinct tests. Separate LOINC names and codes are given to tests that measure the same construct but have different clinical meanings or purposes. Likewise, distinct LOINC codes are assigned for tests whose reference ranges are sufficiently different that they change the meaning of the test. All of the information necessary to disambiguate these tests is supposed to be stored in the six main fields.<sup>21</sup>

Table 1 shows how these six main fields have been adapted for standardized assessment measures. The *Component* field names the construct being assessed; *Property* roughly distinguishes between subjective and objective data; *Timing* indicates whether the measurement refers to a point in time or an interval; *System* indicates the source of data; *Scale* specifies the

type of data collected; and *Method* is supposed to give sufficient information to distinguish among similar versions of the test. Each distinct item (question) within a standardized assessment is assigned a unique LOINC code.

Table 2 shows the supplemental information that may optionally be stored for each item in a standardized assessment instrument. The *Class* field includes the name of the standardized instrument. The *Survey Question Text* field stores the text of the question; and the *Survey Question Source* field specifies the name of the instrument and the question number within that instrument. Some LOINC entries also contain *Formulas*, which are nonstandardized descriptions of the equation used to calculate the LOINC entry. A few LOINC entries also have entries in the *Answer List* column, which stores the allowable values for nominal and ordinal lists. Currently, these supplemental fields are sparsely populated. They provide additional information, but are not considered crucial for distinguishing among related items.

This six-axis schema would be adequate if there were only a single version of each instrument. Unfortunately, some instruments have undergone several revisions, or are locally modified. For example, they might have differing *Answer Lists*, with clearly differing meanings, and thus cannot be compared. In general, any change in the operational definition (*Survey Question Text*) or variable definition (*Answer List*) might change the meaning of the test. Although the LOINC naming principles require that separate codes be assigned in these circumstances, the current six-axis schema is not sufficiently flexible to indicate the presence of variant operational or variable definitions; and the supplemental tables are inadequately standardized to indicate how the versions differ.

Table 1 ■

The Six Main Fields of the LOINC Semantic Model, as Extended for Standardized Nursing Assessment Instruments

Element	Size (chars)	Definition
Component	150	Name of the construct assessed
Property	30	Whether finding or impression
Timing	15	Whether point or interval
System	100	The body or social system assessed (e.g., eye, mouth, patient, family, caregiver, community)
Scale	30	The type of scale (e.g., quantitative, ordinal, nominal, narrative)
Method	50	Whether observed or reported, plus the instrument name (e.g., Observed.Omaha)

Table 2 ■

LOINC Database Fields that Can Contain Supplemental Information for Assessment Instruments

Element	Size (chars)	Definition
Class	20	The name of the instrument (e.g., Nurse.Survey.Omaha)
Survey Question Text	255	The exact text of the question
Survey Question Source	50	[Instrument Name].[Question Number or Reference] (e.g., OMAHA III.26)
Comments	254	Description of construct assessed
Formula	255	Equations for calculated values
Answer List	64K	The list of allowable answers for nominal and ordinal scales

DIALOGIX

Dialogix, developed at Columbia University, is a framework for implementing logical or fully deterministic dialogs between humans and computers.<sup>24-28</sup> The authors developed Dialogix to support the broad range of instruments used within psychiatry, including questionnaires, structured interviews, diagnostic algorithms, and decision trees. The Dialogix framework includes a schema and syntax for defining instruments, an engine for deploying them, and a schema for storing the results and timing information collected when an instrument is used. All instruments are deployed as series of web pages that are tailored to the needs of each subject.

Instruments used in human-computer dialogs vary in the types of data they collect, and the degree to which they require complex validation, calculations, branching, and tailoring of the messages. *Validation* criteria are used to check responses against expectations. They might include, for example, ranges of allowable values (if numeric), or lists of valid choices (if nominal or ordinal). Some instruments require cross-item *validation* criteria. For example, demographic questions asking how long an informant has known the subject must ensure that the response is not larger than the subject's age. Other instruments require runtime *calculations* and scoring. For example, the DISC (Diagnostic Interview Schedule for Children)<sup>29</sup> uses complex algorithms in the course of a given interview to screen for psychiatric diagnoses. *Branching* is the process of determining which questions should be asked. For example in the DISC, only those patients who screen positive for depression are asked detailed follow-up questions. Finally, *tailoring* specifies which words should be used to ask a question. Many questions in the DISC are tailored to refer to prior answers, such as the name and gender of the child, and lists of symptoms and life events.

Dialogix provides an instrument definition schema that fully and explicitly models the conceptual, oper-

ational, and variable definitions for each measurable entity. This schema (Figures 1 and 2) consists of nodes (conceived as *Rationale-Action* or *Why-How* pairs) arranged in a simple, linear, sequence. Each node models either a distinct construct to be measured (e.g., via a question or formula) or a message to be communicated. Within the Dialogix framework, each of these are measurable entities, since Dialogix records not only the answers to questions, but also the amount of time that subjects spend reading instructions and the time and method they use to enter their answers. Whole instruments are defined in a single spreadsheet or table, with one node per row. The *Why*

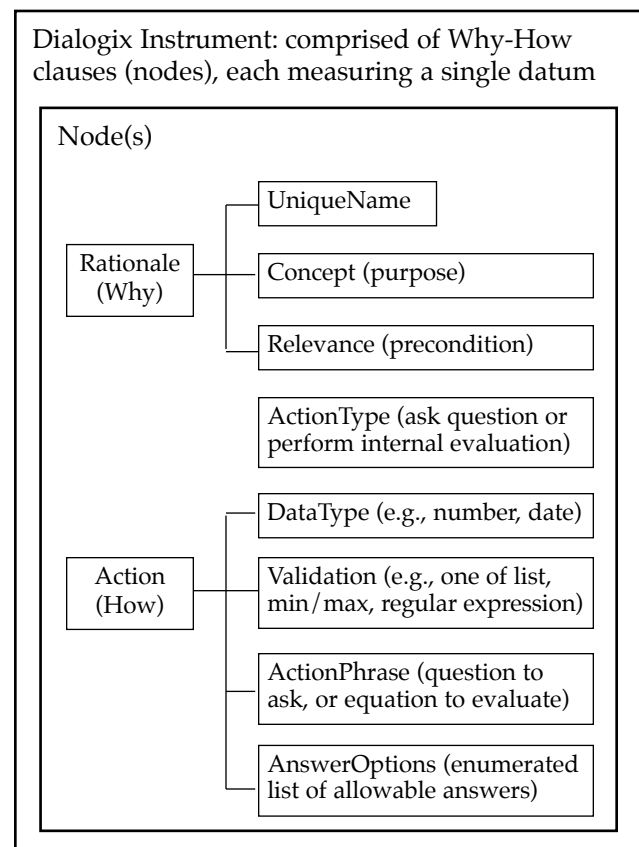


Figure 1 Dialogix schema.

Column Name	Example 1	Example 2	Example 3																																													
Primary Key	123456	123457	123458																																													
ParentLOINC	xxxxx-x	xxxxx-x	xxxxx-x																																													
InstrumentFK	[BPRS-original]	[BPRS-A]	[BPRS-A]																																													
UniqueName	Suspiciousness	Suspiciousness	Andp																																													
Concept	Suspiciousness	Suspiciousness	Anxiety-Depression Subscale																																													
Relevance	1	1	1																																													
ActionType	q	q	e																																													
DataType	number	number	number																																													
Validation																																																
ActionPhrase	Belief (delusional or otherwise) that others have now, or have had in the past, malicious or discriminatory intent towards the patient. On the basis of verbal report, rate only those suspicions that are currently held, whether they concern past or present circumstances.	<b>ASK:</b> "How did you get along with people, in general, during the past week? Do you feel that you have to be on guard with people? Has anyone been giving you a hard time, or accusing you of things? Has anyone deliberately tried to annoy you? Tried to harm you?" <b>RATE VERBAL REPORT OF:</b> Current belief (delusional or otherwise) concerning past or present circumstances.	(DepressiveMood + Anxiety + SomaticConcerns + GuiltFeelings)																																													
AnswerOptions	<table border="1"> <thead> <tr> <th>Code</th> <th>Label</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>Not assessed</td> </tr> <tr> <td>1</td> <td>None</td> </tr> <tr> <td>2</td> <td>Very Mild</td> </tr> <tr> <td>3</td> <td>Mild</td> </tr> <tr> <td>4</td> <td>Moderate</td> </tr> <tr> <td>5</td> <td>Moderately Severe</td> </tr> <tr> <td>6</td> <td>Severe</td> </tr> <tr> <td>7</td> <td>Very Severe</td> </tr> </tbody> </table>	Code	Label	0	Not assessed	1	None	2	Very Mild	3	Mild	4	Moderate	5	Moderately Severe	6	Severe	7	Very Severe	<table border="1"> <thead> <tr> <th>Code</th> <th>Label</th> <th>Anchor</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>None</td> <td></td> </tr> <tr> <td>2</td> <td>Very Mild</td> <td>rare cases of mistrust that may not be warranted</td> </tr> <tr> <td>3</td> <td>Mild</td> <td>occasional cases of suspiciousness that are definitely not warranted</td> </tr> <tr> <td>4</td> <td>Mod</td> <td>frequent suspiciousness OR transient ideas of reference</td> </tr> <tr> <td>5</td> <td>Mod Severe</td> <td>pervasive suspiciousness OR frequent ideas of reference</td> </tr> <tr> <td>6</td> <td>Severe (+delusion)</td> <td>encapsulated delusion(s) of reference or persecution</td> </tr> <tr> <td>7</td> <td>Very Severe (+delusion)</td> <td>pervasive or more widespread, frequent, or intense delusion(s)</td> </tr> <tr> <td>0</td> <td>N/A</td> <td>Not assessed</td> </tr> </tbody> </table>	Code	Label	Anchor	1	None		2	Very Mild	rare cases of mistrust that may not be warranted	3	Mild	occasional cases of suspiciousness that are definitely not warranted	4	Mod	frequent suspiciousness OR transient ideas of reference	5	Mod Severe	pervasive suspiciousness OR frequent ideas of reference	6	Severe (+delusion)	encapsulated delusion(s) of reference or persecution	7	Very Severe (+delusion)	pervasive or more widespread, frequent, or intense delusion(s)	0	N/A	Not assessed	
Code	Label																																															
0	Not assessed																																															
1	None																																															
2	Very Mild																																															
3	Mild																																															
4	Moderate																																															
5	Moderately Severe																																															
6	Severe																																															
7	Very Severe																																															
Code	Label	Anchor																																														
1	None																																															
2	Very Mild	rare cases of mistrust that may not be warranted																																														
3	Mild	occasional cases of suspiciousness that are definitely not warranted																																														
4	Mod	frequent suspiciousness OR transient ideas of reference																																														
5	Mod Severe	pervasive suspiciousness OR frequent ideas of reference																																														
6	Severe (+delusion)	encapsulated delusion(s) of reference or persecution																																														
7	Very Severe (+delusion)	pervasive or more widespread, frequent, or intense delusion(s)																																														
0	N/A	Not assessed																																														

**Figure 2** Three sample entries in the Assessment\_Instrument table. The column names are adopted from the Dialogix schema in Figure 1, with the addition of *ParentLOINC* which links this table to the main LOINC table, and *InstrumentFK* which refers to the LOINC PanelElement that codifies the ordering of questions within an instrument. Note that *UniqueName* must be unique only within the context of a single instrument, thus ensuring that formulas, like that in Example 3, reference the proper variables. Examples 1 and 2 show two different ways of assessing Suspiciousness; and Example 3 shows how formulas are represented using this schema. A *Relevance* of 1 means that the question is always asked. The *ActionTypes* 'q' and 'e' indicate that a question should be asked, and a formula should be evaluated, respectively. None of these items required validation criteria, so that field is blank.

component of a node includes the unique identifier for the variable within the scope of the instrument (*UniqueName*); the conceptual definition (*Concept*), which describes which construct is being measured by a question or communicated by instructional messages; and the *Relevance*, which indicates *when* (the conditions under which) the action should be per-

formed. This *Relevance* field declaratively codifies all of the branching rules in a single Boolean equation and can support complex conditional follow-up, looping, and backtracking to change answers. The *How* component of a node specifies the type of action to be performed (*ActionType*) and has a different syntax depending on whether the action is to ask a question

or perform a calculation. For calculations, the *ActionPhrase* is the formula to be processed and can support logical, statistical, aggregation, and textual operations. For nodes that define questions, the *ActionPhrase* contains the text of the question and thus constitutes the operational definition of the variable. The other fields codify the variable definition for questions by specifying the *DataType*, *Validation* criteria, and *AnswerOptions*, as needed. The *AnswerOptions* field meets this need by specifying the label for each allowable answer, the associated internal code to be used for each, and an optional anchor clarifying the meaning of the label. For example, if the label is "mild" and the question is about the degree of suspiciousness, then the anchor describes what "mild" means in that context (Figure 3). The *AnswerOptions* field also codifies which data input style was used. The list of possible input styles is described in the section on presentation as a source of measurement error.

The Dialogix instrument definition schema supports tailoring of questions and messages by letting the *ActionPhrase* and *AnswerOptions* text fields include embedded tailoring logic. The deployment system parses these fields to resolve the tailoring logic, which can include conditional substitutions (e.g., different words to refer to a children or youths, depending on their age) and locale-specific formatting. These messages can also include embedded Hypertext Markup (HTML) so that key points can be appropriately emphasized. Thus, the instrument definition explicitly models all possible ways of customizing the order (branching) and wording (tailoring) of questions within an instrument.

During deployment of an instrument, the Dialogix engine always starts at the first node in the instrument definition, and sequentially processes the nodes from first to last. For each node, Dialogix assesses whether the node is relevant by processing the *Relevance* field. If the node is relevant, then its action is performed, and the subject is asked a question or given an instructional message. If the node is not relevant, the value of "not applicable" is stored for that variable.

Dialogix deploys all of these instruments as a series of tailored web pages—a human-computer dialog. Each page contains instructions and/or questions with associated answer options. If the respondent provides valid answers for the questions, she is advanced to the next set of relevant ones. If not, she is given helpful error messages that reflect the validation criteria. If she makes a mistake, like erroneously indicating that she is male, she can backtrack, change her answer, and be given gender-appropriate questions from

thereon. Dialogix performs all of the necessary validation, calculations, branching, and tailoring, rather than forcing the respondent to handle these tasks. Respondents move forward or backward one page at a time, since jumping to arbitrary questions can violate the branching logic.

By these means, Dialogix supports the use of instruments defined and deployed as human-computer dialogs. Each node in an instrument definition contains all of the important information methodologically required to specify a single measurable entity – its conceptual, operational, and variable definitions. Thus, the instrument definition explicitly models all possible ways of customizing the order (branching) and wording (tailoring) of questions within an instrument. In deployment, Dialogix records the actual order in which a subject is asked questions, including a log of backtracking to change answers; the fully tailored text of the questions asked; and the time spent and method used to answer each question.

## Model Formulation Process

As part of the formulation process, it was necessary to identify the ways in which instruments can vary; the subset of those that might significantly affect reliability and validity; and the subset of those that can be realistically codified in a semantic model. Then, the modeling had to distinguish between static factors that are amenable to precoordination and dynamic attributes that should be postcoordinated as an annotation of the value in a message instance.

## Sources of Measurement Error

According to survey research, there are four main sources of error, or variance, when asking health questions<sup>9</sup>: (1) how questions are posed, (2) the method used to ask them, (3) who does the asking, and (4) who does the answering. Differences in any of these can affect the reliability and validity of the answers. How questions are posed includes the wording and language of the questions; the context, or order, in which they are asked; and presentation attributes of the text (such as font size and emphasis), which affect the readability and comprehensibility of the questions. The method used to ask them includes face-to-face interviewing, telephone interviewing, computer assisted interviewing, and self-administration. Since different people and classes of interviewers (e.g., level of clinical training) have different biases, the quality of the data collected can vary

<b>15. Suspiciousness</b>							
<b>ASK:</b> "How did you get along with people, in general, during the past week? Do you feel that you have to be on guard with people? Has anyone been giving you a hard time, or accusing you of things? Has anyone deliberately tried to annoy you? Tried to harm you?" <b>RATE VERBAL REPORT OF:</b> Current belief (delusional or otherwise) concerning past or present circumstances.							
<input type="radio"/> 1) none	<input type="radio"/> 2) very mild	<input checked="" type="radio"/> 3) mild	<input type="radio"/> 4) moderate	<input type="radio"/> 5) mod severe	<input type="radio"/> 6) severe (+del)	<input type="radio"/> 7) very severe (+del)	<input type="radio"/> n/a
rare cases of distrust that may not be warranted	occ. instances of suspiciousness that are definitely not warranted	freq. suspiciousness OR transient ideas of reference	pervasive suspiciousness, OR frequent ideas of reference	encapsulated <b>delusion(s)</b> of reference or persecution	pervasive or more widespread, frequent, or intense <b>delusion(s)</b>	not assessed	

**Figure 3** Recent BPRS variant<sup>2</sup> with anchors from the BPRS-A.<sup>3</sup> The dotted rectangle surrounds an answer's label (mild) and score (3). The dashed rectangle surrounds an answer's anchor, which clarifies what qualifies as mild suspiciousness.

based upon who does the asking. Finally, individuals may differ in their capability and willingness to answer questions truthfully, depending on how well they understand the questions, know the answers, and trust the person (or computer) asking them. Although the last two can introduce significant variability, only the first two are properties of the instrument and methodology themselves.

Given that minor variants in the implementation of instruments can result in significant differences in reliability and validity, how many distinct codes should there be? At one extreme, a unique code could be assigned for each minor variant, but this would reduce sharability of data that might actually be comparable. The other extreme, which is the status quo, is to assume that all variants are identical, but this approach risks type I and II errors if the questions actually have different meanings or reference ranges. The optimal approach is to have the minimum number of codes necessary to disambiguate versions with significantly different clinical meanings or reference ranges. However, this poses a catch-22. The only way to robustly assess whether there should be separate codes for two variants is first to assign each a separate code and then to conduct a measurement study to determine whether their psychometric properties are significantly different.

Fortunately, although there are many ways in which instruments can vary, only a subset of them is likely to cause significant problems. Unfortunately, there are no heuristics that predict the degree to which reliability and validity are affected by such changes; further research is needed to clarify the differences among versions. In the interim, it is prudent to use distinct codes for each version and to store all perti-

nent information about the test, so that measurement studies can be performed to detect potential differences. If questions are found to be essentially identical, then their codes can be merged, and values collected using those versions can be reliably compared. Alternatively, if questions are found to have different meanings or reference ranges, efforts can be made to re-scale the values collected using those questions, lest errors be introduced by attempting to compare tests with different meanings.

### Examples of Unreliability

The Brief Psychiatric Rating Scale (BPRS)<sup>1</sup> provides good examples of how variations in wording can affect the meaning and reference range of items (Figure 4). The BPRS is commonly used to measure affective and psychotic symptoms in psychiatric patients. The original instrument includes 18 items, each measured along a common 7-point Likert scale. The individual items values are seldom used. Instead, the scores for each of the 18 items are added to form the total BPRS score, along with 4 subscale scores. For example, the *Anxiety-Depression* subscale score is the sum of the values for the items measuring *depressed mood, anxiety, somatic concerns, and guilt feelings* (see Example 3 of Figure 2).

In the 40 years since the BPRS was created, dozens of variants have been created and used, but the scores are reported as though the original version had been used. These variants include changes in the wording, order, and presentation. Fundamental research into survey design has shown that changes to any of these three factors can affect the reliability and validity of the collected data.<sup>10</sup> Moreover, some variants have different numbers of subscales.<sup>30,31</sup> Other variants

<b>BRIEF PSYCHIATRIC RATING SCALE (BPRS)</b>	
Please enter the score for the term which best describes the patient's condition	
<b>0</b> = not assessed, <b>1</b> = not present, <b>2</b> = very mild, <b>3</b> = mild, <b>4</b> = moderate, <b>5</b> = moderately severe, <b>6</b> = severe, <b>7</b> = extremely severe	
<p><b>1. SOMATIC CONCERN</b> Degree of concern over present bodily health. Rate the degree to which physical health is perceived as a problem by the patient, whether complaints have a realistic basis or not.</p> <p style="text-align: right;">SCORE <input style="width: 30px; height: 20px;" type="text"/></p> <p><b>2. ANXIETY</b> Worry, fear, or over-concern for present or future. Rate solely on the basis of verbal report of patient's own subjective experiences. Do not infer anxiety from physical signs or from neurotic defense mechanisms.</p> <p style="text-align: right;">SCORE <input style="width: 30px; height: 20px;" type="text"/></p> <p><b>3. EMOTIONAL WITHDRAWAL</b> Deficiency in relating to the interviewer and to the interviewer situation. Rate only the degree to which the patient gives the impression of failing to be in emotional contact with other people in the interview.</p> <p style="text-align: right;">SCORE <input style="width: 30px; height: 20px;" type="text"/></p>	<p><b>10. HOSTILITY</b> Animosity, contempt, belligerence, disdain for other people outside the interview situation. Rate solely on the basis of the verbal report of feelings and actions of the patient toward others; do not infer hostility from neurotic defenses, anxiety, nor somatic complaints. (<i>Rate attitude toward interviewer under "uncooperativeness."</i>)</p> <p style="text-align: right;">SCORE <input style="width: 30px; height: 20px;" type="text"/></p> <p><b>11. SUSPICIOUSNESS</b> Belief (<i>delusional or otherwise</i>) that others have now, or have had in the past, malicious or discriminatory intent toward the patient. On the basis of verbal report, rate only those suspicions which are currently held whether they concern past or present circumstances.</p> <p style="text-align: right;">SCORE <input style="width: 30px; height: 20px;" type="text"/></p>

**Figure 4** Part of the original BPRS.<sup>1</sup>

add items to assess constructs not originally covered by the BPRS, including negative symptoms and side effects.<sup>32</sup> Thus, it may be inappropriate to compare data collected from different versions of the BPRS.

### Wording

Variations in wording can occur at several levels. Instruments may be divided into sections or modules. Each module may have identifying labels describing the constructs measured in that section as well as instructions that apply for all of the questions in the module. Individual questions may also have labels identifying or summarizing the construct to be assessed, as well as question-specific instructions. When the scale is nominal or ordinal, the individual answer choices may be anchored with labels and instructions. Nominal and ordinal scales usually also have numerical *coding* values associated with each choice.

For example, in the original version of the BPRS (Figure 4, and Example 1 of Figure 2), question 11 assesses the construct "Suspiciousness." The question's instructions are:

Belief (*delusional or otherwise*) that others have now, or have had in the past, malicious or discriminatory intent towards the patient. On the basis of verbal report, rate only those suspicions that are currently held, whether they concern past or present circumstances.

The ordinal Likert scale is labeled as follows:

0 = not assessed, 1 = not present, 2 = very mild, 3 = mild, 4 = moderate, 5 = moderately severe, 6 = severe, 7 = extremely severe.

Some versions are more explicit. For example, the anchored BPRS (BPRS-A)<sup>3</sup> (Figure 3, and Example 2 of Figure 2) has the following instructions for suspiciousness:

**Ask:** "How did you get along with people, in general, during the past week? Do you feel that you have to be on guard with people? Has anyone been giving you a hard time, or accusing you of things? Has anyone deliberately tried to annoy you? Tried to harm you?"

**Rate verbal report of:** Current belief (*delusional or otherwise*) concerning past or present circumstances.

And the answer choices are explicitly anchored as follows:

- |                      |  |
|----------------------|--|
| 1. None              |  |
| 2. Very mild         | Rare cases of distrust that may not be warranted                         |
| 3. Mild              | Occasional instances of suspiciousness that are definitely not warranted |
| 4. Moderate          | Frequent suspiciousness or transient ideas of reference                  |
| 5. Moderately severe | Pervasive suspiciousness or frequent ideas of reference                  |

- |                              |   |
|------------------------------|---|
| 6. Severe (+ delusions)      | Encapsulated delusion(s) or reference of persecution            |
| 7. Very severe (+ delusions) | Pervasive, or more widespread, frequent, or intense delusion(s) |

More commonly, variations on the BPRS lack labels and anchors for the Likert scale, and many lack question-specific instructions. Thus, typical renditions would simply instruct the clinician to rate “suspiciousness” on a scale from 1 to 7. Moreover, some unpublished, locally modified variants use codes from 0–6 instead of 1–7.<sup>2</sup>

As might be expected, explicitly labeling and anchoring the answer options can dramatically impact reliability and validity,<sup>3</sup> especially in the case of fuzzy ordinal scales like the one used in the BPRS. Without the explicit anchoring, raters may not know that delusions are a requirement for a rating of 6 or 7 on the suspiciousness scale or how well symptoms must be resolved to be labeled as “mild.” The suspiciousness scale is especially problematic at psychiatric prisons, where inmates commonly threaten one another. If the apparent suspiciousness relates to real threats, the subject might get a rating of 1; but a rater unfamiliar with the patient’s environment might give a rating of 7. Thus, the presence and contents of labels, anchors, and codes within the list of allowable answers can distinguish among versions and should be explicitly modeled in LOINC.

Differences in instructions can also alter the reliability and validity. Trained clinicians might not need example questions to elicit suspiciousness. Unfortunately, clinicians often forget that 13 of the 18 items on the BPRS include explicit exclusion criteria within the instructions.<sup>2</sup> For example, on the question of “Conceptual Disorganization,”<sup>2</sup> the instructions are as follows:

**RATE OBSERVED:** Formal thought disorder: loose associations, flight of ideas, incoherence, neologisms

**NOT:** Mere circumstantiality or pressured speech, even if marked; NOR patient’s subjective impression

Without such explicit instructions, clinicians may measure a different construct that includes circumstantiality and subjective impression. It is not uncommon for raters using the less explicit versions of the BPRS to deviate several points from the proper answer on standardized vignettes,<sup>2</sup> thus potentially altering treatment decisions. Thus, the complete instructions for each question must be explicitly modeled.

The situation gets even more complicated with some structured interviews that tailor the content of the questions and answers based on prior answers. For example, the DISC<sup>29</sup> asks which adults have cared for a child in the past year; the two dozen choices include extended relatives, friends, and guardians. Then the DISC asks, “Which of these adults (*list of caretakers*) do you feel closest to?” Interviewers must know that instead of saying “*list of caretakers*,” they should remember which caretakers the subject mentioned and construct a grammatically correct list. Likewise, other questions refer back to gender, demographic information, lists of events, and counts of symptoms. This approach is problematic, because interviewers may forget the prior answers and construct grammatically incorrect or awkward questions. Such errors could damage the rapport between the subject and the interviewer and erode the reliability of the collected data.

Computerization can remove the variability caused by poor memory or grammar, but only if the system supports the accurate retrieval of dependent information and the robust and dynamic composition of sentences. Dialogix, with its embedded tailoring syntax, supports this functionality. Thus, LOINC needs to be able to distinguish among versions that support tailoring and those that do not.

### Ordering

Psychometric research has shown that the order in which questions are asked can affect the results,<sup>14</sup> especially in long, self-administered instruments, during which subjects’ attention might wane. Lengthy instruments or those that are hard to understand can degrade motivation, leading to false, random, or inconsistent answers. Recently asked questions can also affect expectations and alter how subjects respond to questions. Finally, there are well-recognized response styles, such as social desirability, that can affect how honestly subjects answer items. Therefore, the order in which questions are asked can affect reliability and validity.

A distinction should be made between the order in which questions are supposed to be asked and the order in which they are actually answered. The prescribed order is an attribute of the instrument itself. However, the order in which the question are answered is a property of the subject, since branching criteria determine which questions a subject will see, and subjects may be allowed to backtrack and change answers. Ideally, the semantic schema should model both.

## Presentation

Presentation includes both the look and feel of the instructions and questions and also to the manner in which the subject is asked to answer the questions.

The format of text messages can affect how questions and instructions are interpreted.<sup>20</sup> Bolding, italics, colors, fonts, and capitalization can affect the readability of items and change their phrasing. These can also draw attention towards or away from the key parts of the instructions. Thus, such textual formatting can affect reliability and validity. On the other hand, browsers for the visually impaired are able to indicate changes in formatting only by altering the emphasis of the spoken text. In such cases, only the location of the emphasis is likely to matter.

The means by which subjects indicate their answers can also affect the results.<sup>9</sup> There are a limited number of typical input styles for data, depending on the type of data being collected. Textual data are normally entered through typing or voice transcription. Nominal and ordinal data require selection among several categories, which can be represented visually by using graphical input elements such as radio buttons, check boxes, list boxes, drop-down boxes, and text fields. Quantitative data are often entered using free text fields or graphical sliders. Some of these approaches make all allowable choices visible, whereas others, such as the drop-down box, may not. Thus, different input styles may impose different cognitive loads, as well as require different levels of training, either one of which can impact the accuracy of the data collected.

In summary, since changes in wording, order, and presentation can affect the meaning and reference range of items within standard instruments, LOINC should model enough of these attributes to distinguish among versions of instruments. Specifically, the wording and presentation features that codify the operational and variable definitions should be modeled. Moreover, since the ordering of questions within an instrument can affect reliability and validity, the expected ordering (as defined by the instrument), and the actual ordering (as used by a subject, who might backtrack to change answers) should be captured.

### Modeling Static and Dynamic Aspects of Measurement Error

The four sources of measurement error described by Aday<sup>9</sup> include a combination of static and dynamic

factors. The static aspects comprise the instrument definition, and should be referenced by precoordinated LOINC names, whereas the dynamic aspects comprise the message instance and should be post-coordinated as an annotation to the value.

How questions are posed and the methods used to ask them consist of a mixture of static and dynamic factors. The static factors include the untailed text of the question, any embedded presentation attributes, and the default or expected order in which the question should be asked. Essentially, these static factors are the operational and variable definitions of the instrument definition and specify all of the possible ways in which the questions might be asked. The dynamic, or subject-specific, attributes of a question include the language used, the fully tailored text, the sequence in which it is asked, the modality used to ask it (such as face-to-face, telephone, or self-administered<sup>9</sup>), the speed with which it is answered, and possibly the keystrokes or mouse movements used to enter the answer.

Static properties are amenable to modeling using precoordinated LOINC names. A static instrument definition can be created for each variant of an instrument, even those that support complex branching and tailoring. The LOINC Method element can be extended to reference that static definition, whose details would be stored in a supplemental table. Thus, the operational and variable definitions do not need to be passed as part of the LOINC name.

In contrast, the dynamic factors need to be post-coordinated as contextual attributes of the value part of the name-value pair. For example, the reference range of an assessment item's value is effectively defined by the text of the variable definition; thus the reference range will be dynamic if any tailoring is done. The HL7 data exchange standard supports the transmission of reference ranges as an attribute of the name-value pair in the observation reporting (OBX) segment. Although HL7 can transmit textual reference ranges, the text of the variable definition can exceed the maximum length allowed by HL7. One convenient alternative might be to transmit and post-coordinate this information using the annotation (ANO) field within the OBX segment of an HL7 message. Since Dialogix captures and stores this dynamic data, its schema could be modified for use by the ANO field. However, the modeling of these dynamic features within data exchange standards is beyond the scope of this article.

## Model Description

We propose four modifications to the LOINC naming schema. First, a separate table, entitled *Assessment\_Instruments*, should be created, containing the eight fields of the Dialogix schema (see Figures 1 and 2). This table should also include a *ParentLOINC* field that identifies the associated LOINC code for the question and an *InstrumentFK* field that points to the LOINC Panel that defines the ordering of questions within an instrument. This approach allows the supplemental information needed for instruments to be fully stored without cluttering the schema of the LOINC table. Second, the *Survey Question Text*, and *Survey Question Source* fields can be removed from the main LOINC table, since these are better modeled within the *Assessment\_Instruments* table. Third, each distinct instrument should be stored as a separate LOINC Panel. Fourth, the naming schema within the Method section should be refined to use the following syntax: [Observed or Reported].[Instrument Name].[Variant Identifier], where Variant Identifier is a non-semantic index of identified variants in the wording, order, or presentation of that question.

The proposed model is an outgrowth of that used for the Dialogix system. As previously described, Dialogix already provides a schema for the conceptual, operational, and variable definitions of each measurable entity. Thus, it meets the needs for modeling the differences in wording and presentation. Moreover, Dialogix meets the needs to model ordering, even for instruments that include complex branching. Thus, the Dialogix instrument definition schema can formally model questions from assessment instruments.

The *Assessment\_Instruments* table would contain all of the information codified in the Dialogix schema. Each of these new columns would have a distinct XML (extensible markup language)<sup>33</sup> syntax that conforms to the Dialogix model for those fields and maximizes the generalizability of LOINC by using Unicode<sup>34</sup> and standard language identification codes.<sup>35</sup> Figure 2 shows how the *Assessment\_Instruments* table would represent the BPRS Suspiciousness questions, and the Anxiety-Depression subscale calculation.

If the eight columns of the Dialogix schema were added to LOINC to support this functionality, there could be conflicts with the existing survey-specific extensions to LOINC. As seen in Table 2, the existing LOINC extensions provide storage for the text of the

questions in the *Survey Question Text* field, the list of allowable answers in the *Answer List* field, and an indication of the order of the question within the instrument in the *Survey Question Source* field. Moreover, the *Comments* field can store extra information about the conceptual definition. Finally, the *Formula* field can store equations, such as those needed to calculate the BPRS subscores.

These existing LOINC fields would need additional modifications to meet their currently specified purpose. The *Survey Question Text* field is limited to 255 characters, which is not enough to store many of the lengthy questions found in instruments like the DISC.<sup>29</sup> The *Answer List* field contains plenty of room to store the possible answers for nominal and ordinal scales, but it does not specify a consistent syntax that specifies the internal codes or anchors associated with the labeled answers. The *Survey Question Source* and *Comments* fields may be of adequate length to provide an indication of the location of the question within the instrument and a formal concept definition, respectively. However, the *Formula* field is not sufficiently long to store equations such as those needed for calculating the total BPRS score; and the field does not currently have a formal syntax. Therefore, these existing supplemental columns within LOINC need to be modified or replaced.

Since only 460 assessment instrument elements have been added to LOINC as of the July 2001 release, a potentially viable alternative is to migrate those instrument elements to the Dialogix schema and remove the redundant LOINC fields. Specifically, the *Survey Question Source* and *Survey Question Text* columns can be removed from LOINC. Instead, the *Survey Question Source* information would be stored in the *Assessment\_Instrument* table, and LOINC PanelElement, as described below. The content of the *Survey Question Text* field can be replaced by the *ActionPhrase* field, which provides a consistent syntax for tailoring. The same *ActionPhrase* field also provides a standard syntax for modeling equations, so that it can replace and enhance the LOINC *Formula* field. The LOINC *Answer List* field can be fully modeled within the Dialogix *AnswerOptions* field, with the added benefit of a codified schema for indicating the labels, anchors, and codes. Finally, rather than use the LOINC *Comment* field to indicate the conceptual definition of a variable, the Dialogix *Concept* field could be used.

To define instruments completely, the order of questions within them and the branching logic also needs to be specified. LOINC supports a feature called pan-

els, which are collections of tests. The tests that comprise the panel are stored in the *PanelElements* field as a set of semicolon delimited names. These names are not unambiguous references to other LOINC codes; thus the field is informative rather than definitional. The LOINC panel feature could be extended to support assessment instruments. Each complete instrument could be stored as a panel, with the *PanelElements* field being a list of references to the LOINC codes that define the nodes within the instrument. We propose that each test variant be assigned a unique LOINC panel. This would allow the disambiguation of instruments with similar questions, but different default orders. Moreover, additional metadata describing the instrument could also be stored in the panel element, thereby replacing and enhancing the *Survey Question Source* field. For example, the name, version, authors, date, and purpose of each instrument should be stored as metadata. Appropriate finite sets of attributes can be adapted from those used in Arden syntax modules<sup>36</sup> and GLIF guidelines.<sup>37</sup> Alternatively, a separate *Instrument\_References* table could be created to store this information.

The method element would need to be modified to reference the possible variants of the instrument. Rather than trying to describe the differences between the methods in a few characters allowed within the method field, we propose that a non-semantic identifier, like an incremental counter, be used to serve as the *Variant Identifier*. The details of the operational and variable definitions would be stored in the associated row of the *Assessment\_Instrument* table. This approach results in only a minor modification of the method element, while still disambiguating among versions.

When there is only one known variant of each question in an instrument, a single code per question is adequate. In contrast, when there are N known variants, (N+1) codes would be needed, assembled as an implicit, two-level hierarchy. Each known variant would have its own code, which would represent the formal description of the methods contained within the *Assessment\_Instrument* table. In addition, a separate, root-level, code would be needed to indicate that the methods are unknown. The root-level code would not specify the *Variant Identifier* and would be the parent of each of the more specific methods. This would allow data from all variants to continue to be aggregated, as they are now, but would also allow data from identifiable versions to be separately analyzed.

Making the suggested modifications to LOINC would be backward compatible and support considerable future growth. The 460 existing codes for nursing assessment instruments could easily be migrated to the Dialogix schema. The new schema would support a much broader range of assessment instrument types.

## Validation

The Dialogix schema has been used to computerize and deploy a wide variety of instruments. These include two large epidemiological trials,<sup>38,39</sup> each with nearly 2000 highly branched, multilingual, tailored questions. As of March 2002, Dialogix has been used by over 50 interviewers to collect data from over 3000 subjects as part of these epidemiological studies. Dialogix has also been used to web-enable decision trees, clinical guidelines, anonymous surveys,<sup>27</sup> and consumer-oriented decision support tools.<sup>40</sup> We have also implemented several dozen psychiatric instruments,<sup>24</sup> and have validated that Dialogix can implement all of the psychiatric instruments that are not copyright protected.<sup>26</sup>

These prior results, showing the range of instruments that can be web-enabled using Dialogix, affirm that the proposed extensions to LOINC can operationalize the required instruments and distinguish among their various versions.

## Discussion

This schema can resolve two limitations in the existing survey-specific extensions to LOINC. First, it can support complex instruments, such as those requiring branching, calculations, or tailoring. Second, it can distinguish among versions of instruments whose changes in wording, order, or presentation can affect the reliability or validity of the instruments.

Moreover, by adopting these recommendations, LOINC will contain fully operationalizable versions of assessment instruments. Linked with a Dialogix-like engine, users could search for the panel that defines the instrument, and a parser could dynamically extract the component questions, instantiate the instrument, and ensure that the collected data is LOINC-compatible.

The ability to completely formalize and represent standardized instruments and questions may encourage more researchers to evaluate the quality of these instruments. Currently, researchers must rely on pub-

lished reliability and validity studies, even though many are quite outdated. Moreover, the most commonly used instruments are copyright protected, or sold for a profit, so it is difficult and expensive to cross-compare instruments. Despite the fact that there are guidelines for writing quality questions (e.g., avoiding double negatives, double barreled questions, vagueness),<sup>10</sup> these guidelines have been routinely violated in many of the established instruments, resulting in a clamor for new and improved assessment instruments.<sup>17-19</sup> Making it easier to represent and classify questions may lower the barriers to creating and validating better instruments.

Moreover, this approach could facilitate the efforts of journals that want the survey methodologies to be better clarified within the Methods section of publications. Some epidemiology journals request that copies of the assessment instruments be linked to their web pages.<sup>18,19</sup> Using this LOINC-based approach, fully implementable versions of the instruments could be linked to the web site and described in the Methods section, thus facilitating the analysis and scrutiny of those methods.

Finally, the availability of better-defined instruments can help efforts to detect and reduce medical errors. Such an approach would facilitate the measurement studies needed to identify optimal instruments. Moreover, storing data using codes that disambiguate versions would allow for retrospective review and mining of data for possible errors. For example, the time required to read and answer questions, and details about the path that a subject takes through an instrument, can be used to identify systematic problems in the wording or presentation of questions.<sup>40</sup> This information can also be used to detect changes in the reliability and validity of instruments as they are deployed in new settings, populations, or languages.

One potential limitation is that the proposed recommendations would result in the generation of many new LOINC codes. For example, the authors have seen nearly two dozen unpublished BPRS variants, although an exact count would require a formal study. In the short term, this may seem more work than is cost effective, helpful, or reasonable. Although there would be many unique codes, there would also be hierarchies of related codes, thus allowing aggregation of data from different versions of instruments until measurement studies can be performed to assess whether items really measure the same concepts with similar sensitivity and specificity. Unfortunately, it may be many years before measurement studies are

conducted that clarify the differences among the various versions of individual items. Thus, unless a schema like that proposed is used to represent and collect the data, it will be impossible to conduct these measurement studies, and a range of possible medical errors may go undetected.

One could also argue that storing tailoring and branching criteria within LOINC violates its purpose. Likewise, the proposed modification to the PanelElement could be questioned. LOINC traditionally classifies collected data, not presented information. However, since tailoring can alter the meaning and reference range of variables, it is appropriate for LOINC to distinguish among variants that have differing tailoring criteria. Moreover, since classification schemata for presented information are lacking, LOINC may be a reasonable place to start.

If such an approach is adopted, additional research is needed to model the metadata required to robustly describe instruments. Further research and modeling also are needed to describe the schema for associating context with values in a postcoordinated fashion; and for identifying the amount of detail that might be valuable for measurement studies.

These proposed extensions to the LOINC naming schema have a great potential to facilitate the computerization of instruments and their linkage to electronic medical repositories. They might also evolve into a standard for representing assessment instruments.

## Conclusion

The laudable goal of creating a standard representation of assessment instruments to support sharing and re-use can be hampered by differences in implementations that affect the reliability and validity of the measured constructs. This article proposes four extensions to the LOINC semantic schema to support the disambiguation of versions of instruments along the axes of wording, question ordering, and presentation styles: (1) use the Dialogix schema to codify the definition of assessment instruments and the variables they measure, (2) remove the recent survey-specific extensions to LOINC, (3) extend the meaning of LOINC Panels to support the definition of entire instruments, and (4) extend the syntax of the LOINC Method element to distinguish among versions of and acquisition modes for standard instruments. The feasibility of this model is affirmed by the fact that it is a subset of one used by an existing tool that has

implemented a broad range of psychiatric instruments. Such extensions could facilitate the process of conducting measurement studies. In addition, this model could help reduce medical errors caused by comparing data from unreliable instruments.

The authors thank their advisors, Drs. Vimla Patel and Stephen Johnson; Gerry Segal for his mentorship; Dr. Molly Finnerty for introducing them to the needs of assessment instruments; Drs. Hector Bird, Patricia Cohen, Christina Hoven, and Michael Terman for their mentorship and collaboration; and Drs. Jim Cimino and Suzanne Bakken for their many valuable critiques of previous versions of this manuscript.

### References ■

- Overall JE, Gorham DR. The Brief Psychiatric Rating Scale. *Psychol Rep.* 1962;10:799–812.
- Finnerty M, White TM, Buscema C, Lauve T. Clinical practice guideline implementation: Delivery of evidence-based care. In NYS Office of Mental Health 12th Annual Research Conference, 1999 Dec. 6–8, 1999.
- Woerner MG, Mannuzza S, Kane JM. Anchoring the BPRS: an aid to improved reliability. *Psychopharmacol Bull.* 1988;24(1):112–7.
- Brown PJ, Price C. Semantic based concept differential retrieval and equivalence detection in clinical terms version 3 (Read Codes). *Proc AMIA Symp.* 1999:27–31.
- Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. *J Am Med Inform Assoc.* 2000;7(3):288–97.
- Bakken S, Cimino JJ, Haskell R, et al. Evaluation of the clinical LOINC (Logical Observation Identifiers, Names, and Codes) semantic structure as a terminology model for standardized assessment measures. *J Am Med Inform Assoc.* 2000;7(6):529–38.
- Czaja R, Blair J. *Designing Surveys: A Guide to Decisions and Procedures.* Thousand Oaks, CA: Pine Forge Press, 1996.
- Babbie ER. *Survey research methods*, 2nd ed. Belmont, CA: Wadsworth, 1990.
- Aday LA. *Designing and Conducting Health Surveys: A Comprehensive Guide*, 2nd ed. San Francisco, Jossey-Bass, 1996.
- Rossi PH, Wright JD, Anderson AB. *Handbook of Survey Research.* New York, Academic Press, 1983.
- Sudman S, Bradburn NM. *Asking Questions.* San Francisco, Jossey-Bass, 1982.
- Fowler FJ. *Improving Survey Questions: Design and Evaluation.* Thousand Oaks, CA, Sage Publications, 1995.
- DeVellis RF. *Scale Development: Theory and Applications.* London, Sage Publications, 1991.
- Nunnally JC, Bernstein IH. *Psychometric Theory*, 3rd ed. New York, McGraw-Hill, 1994.
- Stone DH. Design a questionnaire. *BMJ.* 1993;307(6914):1264–6.
- Murray P. Fundamental issues in questionnaire design. *Accid Emerg Nurs.* 1999;7(3):148–53.
- Eaden J, Mayberry MK, Mayberry JF. Questionnaires: The use and abuse of social survey methods in medical research. *Postgrad Med J.* 1999;75(885):397–400.
- Olsen J. Epidemiology deserves better questionnaires. IEA European Questionnaire Group. *International Epidemiological Association. Int J Epidemiol.* 1998;27(6):935.
- Wilcox AJ. The quest for better questionnaires. *Am J Epidemiol.* 1999;150(12):1261–2.
- Wyatt J. Same information, different decisions: Format counts. Format as well as content matters in clinical information. *BMJ.* 1999;318(7197):1501–2.
- Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, et al. Logical Observation Identifier Names and Codes (LOINC) database: A public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem.* 1996;42(1):81–90.
- Huff SM, Rocha RA, McDonald CJ, et al. Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. *J Am Med Inform Assoc.* 1998;5(3):276–92.
- Dolin RH, Alschuler L, Beebe C, et al. The HL7 Clinical Document Architecture. *J Am Med Inform Assoc.* 2001;8(6):552–69.
- White TM, Hauan MJ. The capture and use of detailed process information in the dialogix system for structured web-based interactions. *Proc AMIA Symp.* 2001:761–5.
- White TM, Hauan MJ. Dialogix: a System for Rapidly Developing, Deploying, and Analyzing Research Studies. <<http://www.dianexus.org:8080/>>, 2001.
- White TM, Hauan MJ. Dialogix: A framework for improving the accessibility and quality of measurement instruments. In: 15th Annual NIMH conference on Mental Health Services Research, April 1, 2002, Washington, DC, 2002.
- Hauan MJ, Patrick TB, White TM. The individual perspective on patient-identifiable health information in a large midwestern university. *Proc AMIA Symp* [submitted].
- Hauan MJ, White TM, Johnson SB. Dialogix: a system to computerize structured instruments [submitted].
- Shaffer D, Fisher P, Lucas C, NIMH DISC. In Editorial Board: NIMH DISC-IV: Diagnostic Interview Schedule for Children. Parent Informant (Interview about Child)—Epidemiologic Version Edition. New York, Columbia University, 1997.
- Burger GK, Calsyn RJ, Morse GA, et al. Factor structure of the expanded Brief Psychiatric Rating Scale. *J Clin Psychol.* 1997;53(5):451–4.
- Burger GK, Calsyn RJ, Morse GA, Klinkenberg WD. Prototypical profiles of the Brief Psychiatric Rating Scale. *J Pers Assess.* 2000;75(3):373–86.
- Dingemans PM, Linszen DH, Lenior ME, Smeets RM. Component structure of the expanded Brief Psychiatric Rating Scale (BPRS-E). *Psychopharmacology (Berl).* 1995;122(3):263–7.
- Bray T, Paoli J, Sperberg-McQueen CM, Mahler E (eds). *Extensible Markup Language (XML) 1.0*, 2nd ed. World Wide Web Consortium, 2000.
- Unicode Consortium. *The Unicode Standard, Version 3.0.* Reading, MA, Addison-Wesley Developers Press, 2000.
- IETF (Internet Engineering Task Force). RFC 1766: Tags for the Identification of Languages. 1995.
- Hripcsak G, Ludemann P, Pryor TA, et al. Rationale for the Arden Syntax. *Comput Biomed Res.* 1994;27(4):291–324.
- Peleg M, Boxwala AA, Ogunyemi O, et al. GLIF3: the evolution of a guideline representation format. *Proc AMIA Symp.* 2000;(20(Suppl)):645–9.
- Bird HR, Canino G. Boricua Youth Study (BYS). In: *Antisocial Behaviors in US and Island Puerto Rican Youth* (5/1/98–2/8/03), NIMH Grant MH56401, 2000.
- Cohen PR. Children in the community (CIC) study. In *Accounting for Change in Psychopathology—Ages 17–25.* NIMH Grant MH54161-01. 2000.
- White TM, Hauan MJ. Using client-side event logging and path tracing to assess and improve the quality of web-based surveys. *Proc AMIA Symp* [submitted].