Enhancing Biomedical Terminologies to Include Behavioral Health: A Prerequisite to Improving the Quality of Healthcare

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA

BY

Piper Allyn Svensson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Christopher G. Chute, MD, DrPH (Advisor)

May, 2013

Table of Contents

List of Tables

List of Figures

CHAPTER 1
INTRODUCTION

1.1 Introduction

More than a decade ago the Institute of Medicine (IOM) published two landmark reports: *To Err is Human*[1] and *Crossing the Quality Chasm*[2]. These reports highlighted significant deficits in the quality of healthcare delivered in the United States. In 2006, the IOM released the latest report in this series, *Improving the Quality of Care for Mental Health and Substance Use Conditions*[3]. Like earlier reports, this report identified problems with the quality of care delivered in the United States. Unlike earlier reports, however, this report highlighted the importance of including mental health and substance use conditions (MH/SU) in the overall strategy for improving healthcare.

A primary conclusion of the 2006 IOM report is that the framework for quality improvement outlined in earlier reports is as applicable to behavioral health as it is to general healthcare. An overarching recommendation of the report is that the National Health Information Infrastructure (NHII) be expanded to include behavioral healthcare and that health information technology (HIT) be more effectively leveraged to enhance the evidence base for behavioral health conditions. More recently, the United States government enacted legislation in line with the agenda articulated by the IOM. They did this with the HITECH (Health Information Technology for Economic and Clinical Health) Act[4], passed as part of the American Recovery and Reinvestment Act of 2009[5]. This law gives the US Department of Health and Human Services (HHS) the authority to create programs to improve the quality, safety, and efficiency of healthcare in the United States[6]. Embracing the strategy outlined by the IOM, legislators are using this act to promote the use of electronic health records (EHRs) and structured electronic health information exchange (HIE). Regulations created under HITECH define the minimum acceptable standards for electronic health records and specify criteria that must be met to achieve “meaningful use” (i.e., inclusion of features known to improve outcomes) of these systems.

Because those features shown to improve quality (e.g., computerized provider order entry, clinical decision support) depend on the ability of EHRs to code clinical data in a structured and standardized format[7,8,9,10], the ability to achieve these objectives is contingent upon the existence of high-quality terminologies. For behavioral health, this means that existing clinical terminologies must be capable of representing behavioral health concepts at a much more granular level than is currently supported by commonly used coding systems (e.g., the Diagnostic and Statistical Manual of Mental Disorders[11,12] (DSM), International Classification of Diseases[13,14] (ICD), Current Procedure Terminology[15] (CPTS), and the International Classification of Functioning, Disability and Health[16] (ICF)).

The demand for terminologies capable of representing granular behavioral health constructs is not being driven solely by HIT, either. In 2008, the Nation Institute of Mental Health (NIMH) announced a strategic plan emphasizing the need to develop new approaches to classifying psychopathology[17]. These approaches, they argue, should be based on dimensions of observable behavior and neurobiological measures[17] as opposed to categorically conceived syndromes. In 2009, the NIMH established research groups to define core domains of psychosocial functioning and their constituent elements[18]. To date, six specific domains have been identified, and work defining core constructs within each domain is currently underway. The American Psychiatric Association (APA) has similarly eschewed a strictly categorical approach to conceptualizing and diagnosing mental health conditions[19,20]. The upcoming release of the newest version of the DSM (DSM-5)[12], which takes an explicitly dimensional approach to conceptualizing psychopathology, is evidence of the extent to which scientists in this field are emphasizing the importance of granular clinical constructs.

The premise of the current study is that the ability to effectively leverage information technology to improve the quality of care in this domain will require focusing first and foremost on systems for granular concept representation (i.e., clinical terminologies). Specifically, healthcare terminologies must be capable of representing the same breadth and depth of behavioral health constructs as they do medical constructs. At minimum, healthcare terminologies must be capable of representing constructs identified by the NIHM in their specification of the six high priority domains of functioning. They must also be capable of representing the many granular, cross-diagnostic constructs upon which DSM-5 diagnostic categories are based. Finally – and as explicitly argued in the IOM report – terminologies must be capable of representing not only the constructs measured by an extensive inventory of psychological assessment instruments, but also the instrument components and administration procedures associated with these instruments.

The current study was designed to address this need. Focusing specifically on psychological assessment instruments and instrument-related activities, we identify and evaluate terminology systems relevant to the representation and exchange of instrument-related data. We address the extent to which existing terminologies can be used to capture, code, aggregate, and retrieve information in this domain. We selected psychological assessment instruments as the focus of the current study because instruments play a central role in shaping current understanding of clinically relevant phenomena and conditions. Moreover, results obtained using instruments are the grounds upon which clinical practice in this domain is designated as “evidence based”.

In the following sections we present a general overview of the significance of psychological instruments. We then introduce the objectives and specific aims of this study.

1.2 Background

1.2.1 The Significance of Psychological Instruments in Healthcare

Psychological instruments have varying functions depending on which domain in the larger healthcare enterprise they are used. Instruments are both developed and used in psychometric and primary psychological research. They are used in the clinical domain for patient care, outcome evaluation, and the development of evidence-based interventions. Finally, psychological instruments are used in quality measurement and improvement initiatives, including both state and federal accountability programs.

Within each of these domains, a number of core activities are performed relative to psychological instruments. Some of these activities are common to all domains; others are specific to a particular domain. For example, instrument identification is an activity that occurs across the entire healthcare enterprise, from psychometric research, to patient care, to quality improvement. Instrument development, on the other hand, is an activity that occurs primarily within the experimental and clinical research domains.

1.2.2 Instrument Identification

Instrument identification refers to those activities related to the unambiguous identification of a given instrument or instrument component (e.g., instrument scale, subscale, or item). Instrument identification activities include documenting the instrument used in a clinical setting, citing the instrument used in a particular study, and indexing a publication using the name of the instrument discussed or used in the publication. Information elements relevant to instrument identification include those elements that allow a user to successfully differentiate between the instrument of interest and closely related instruments.

1.2.3 Instrument Development

Instrument development includes those activities related to the conceptualization, development, testing, and ongoing refinement of an instrument over time. Information elements relevant to instrument development include those elements that provide insight into the quality of the process used in developing the instrument. This includes elements related to initial item selection, item refinement, and item performance; factor identification and model metrics; and scale and subscale development. It also includes information elements related to the populations and settings in which the instrument was developed, tested and validated. In short, information elements relevant to the instrument development activity include those elements required to describe the life of the instrument from author conception through physical development, refinement, validation and release.

1.2.4 Instrument Evaluation

Instrument evaluation includes those activities related to assessing instrument performance in various contexts following the instrument's initial release. These activities include explicit evaluations of instrument performance, as well as the many implicit evaluations that occur during the routine use of the instrument in research and clinical care. Information elements relevant to these activities include those related to the psychometric properties of an instrument as it is used in-vivo. These elements include sample characteristics, context of administration, and a number of metrics used to describe validity and reliability. Instrument evaluation also includes those activities, and associated information elements, related to formal reviews and evaluations of an instrument.

1.2.5 Instrument Selection

Instrument selection refers to activities related to selecting a specific instrument to meet a specific healthcare objective. This includes selecting an instrument to use in the instrument-development process of another instrument; choosing the appropriate instrument for a specific clinical objective; and selecting the appropriate instrument for a specific research objective. The information elements most relevant to these activities are those which allow the experimental psychologist, clinician, or clinical researcher to determine which of the many available instruments is most appropriate for achieving his or her specific objectives. Examples include elements related to the scientific aspects of the instrument (e.g., construct assessed, populations in which the instrument has been used, and the psychometric properties of the instrument relative to these populations). Other examples include those elements related to the feasibility of using the instrument in a particular context for a particular purpose (e.g., instrument cost, availability, time and training required to administer the instrument).

1.2.6 Instrument Use

Instrument use includes those activities related to the actual use of the instrument in clinical care or research. Information elements relevant to this activity include those related to the context in which the instrument is used. Other relevant elements include details related to administration of the instrument, such as characteristics of the individual administering the instrument, and those of the individual being assessed using the instrument. Finally, the methods and norms used to evaluate instrument results, as well as variables considered in making clinical inferences based on the obtained results, are also relevant information elements.

1.2.7 An Integrated View

None of the individual instrument-related activities occur in isolation from other activities. Moreover, some classes of activities are directly dependent upon other classes. For example, information from all classes of activities are necessary for making evidence-based decisions regarding the appropriateness of an instrument for use in a particular setting, with a particular patient, for a particular goal. Information obtained from all of the above classes of activities is also highly relevant in performing comparative effectiveness research. That is, the quality of the instrument development process, the intrinsic properties of the instrument, and the appropriateness of the instrument for the study population in which it was used all play an important role in assessing the information acquired across a set of studies in a systematic literature review.

1.3 Study Objectives

The primary objective of the current study is to identify and critically evaluate healthcare terminologies relative to psychological assessment instruments and key instrument-related activities. This study focuses specifically on terminologies required to capture, code, aggregate and retrieve information related to psychological assessment instruments. A secondary objective of this study is to present a preliminary information model and controlled vocabulary for instruments and instrument-related activities. These objectives are achieved through three distinct analyses.

The first analysis addresses the need for an explicit information model and controlled vocabulary for psychometric instruments. In this study, we review and harmonize representations of psychological instruments found in current healthcare terminologies, ontologies, and information products. We identify a comprehensive set of 86 instrument-related information elements, and organize these elements into eight classes of information. We empirically assess the form (i.e., explicit, implicit) and consistency with which these representations are implemented. Finally, we integrate these information elements into a coherent model, and explicitly define each of the entities, relationships, and value sets generated from the harmonization process.

The second analysis evaluates the extent to which existing terminologies cover a set of instruments commonly used in behavioral health today. In this study, we evaluate three terminologies most relevant to representing psychometric instruments and other psychological assessments: CDISC's QS Terminology[21,22] (Clinical Data Interchange Standards Consortium Questionnaire Terminology), LOINC[23] (Logical Observations, Identifiers, Names and Codes), and SNOMED CT[24] (Standard Nomenclature of Medicine – Clinical Terms). Using a list of mental health conditions and topics listed on the National Institute of Mental Health (NIMH) web site[25], we assess how well each terminology covers nine discrete classes of information related to instruments.

The third, and final, analysis evaluates the accuracy and completeness of data contained in PsycINFO®. PsycINFO® is the most comprehensive database currently available for accessing information about in-vivo instrument performance. Using a sample of 203 publications, we empirically assess the accuracy and general quality of data in the *Tests and Measures* (TM) field of PsycINFO®.

CHAPTER 2

TOWARD A STANDARD REPRESENTATION OF PSYCHOLOGICAL INSTRUMENTS: A PRELIMINARY INFORMATION MODEL AND CONTROLLED VOCABULARY

In general healthcare, clinicians and researchers have access to a multitude of instruments and technologies for objectively measuring clinical phenomena. In behavioral health, psychological assessment instruments are the primary method for obtaining objective clinical data. Consequently, the quality of these instruments and the appropriateness of their use and interpretation play a critical role in the quality of care delivered in this domain. Given the significance of instruments in a number of healthcare processes, it is essential that information about the intrinsic structure, in-vivo performance, and clinical utility of these instruments be shared, aggregated, and continuously evaluated. To enable the efficient collection, sharing, and aggregation of this data, a robust information model and controlled vocabulary must exist. A prerequisite to developing such a model is the identification and explicit definition of core information elements currently in use across the healthcare enterprise. To date, no such analysis has been performed. The goal of the current study is to fill this gap. Using twelve current industry representations of psychological instruments, we identified, extracted and harmonized a master set of information elements related to instruments. From this analysis, we identified 86 core information elements in eight classes of information. We analyzed the form, frequency, and consistency of representation of these elements across datasources. Our results indicate that information elements included in current healthcare representations vary not only in the type of information they represent, but also in the variety, form, and consistency of representations. Moreover, we found significant gaps in terminologies related to instrument identification, administration and score interpretation activities. We explore the implications of these findings for the development of more robust concepts for instrument representation in existing terminologies. We propose that systematic efforts be made to enhance SNOMED CT and LOINC to provide better coverage of this domain.

2.1 Introduction

In 2006, the Institute of Medicine (IOM) published an extensive report arguing that current efforts to improve the quality of healthcare must be extended to behavioral health. A key recommendation of this report is that the national health information infrastructure (NHII) be enhanced to accommodate the needs of behavioral health[3]. In this report, the IOM specifically identifies the creation of standard terminologies for all commonly used interventions and assessment methods as a prerequisite to building this infrastructure. The current study was designed to address this need. It does so by focusing on terminology requirements related to a large class of assessment-related entities and activities: the use of standardized psychological assessment instruments. In this study, we identify, harmonize, and explicitly define a core set of instrument-related information elements currently in use across the healthcare enterprise. Using the identified elements, we develop a preliminary information model and controlled vocabulary for psychological assessment instruments. Our hope is that the proposed model will be used to stimulate and support future terminology development efforts in this domain.

An information model is generally defined as a formal representation of the concepts, relationships, and operations that specify the semantics for a given domain[26]. That is, an information model explicitly communicates the information requirements for the domain. A controlled vocabulary, on the other hand, defines the finite set of values allowed for a given information element within the model. For purposes of the current study, a psychological assessment instrument is defined as a structured, pre-defined procedure for collecting standardized, quantifiable data related to psychological or behavioral processes. As used in this study, the term ‘psychological assessment instrument’ is synonymous with the term “psychological test”, which is generally defined as “a set of stimuli administered to an individual or a group under standard conditions to obtain a sample of behavior for assessment”[27]. Included in this definition are psychometric instruments, psychological assessments and inventories, rating scales, surveys, questionnaires, and checklists.

The current study extends the work of previous efforts in informatics by viewing psychological instruments from the perspective of the overall healthcare-enterprise and emphasizing representations that support the integration of data from all instrument-related activities across the enterprise. The current study also extends the work of previous efforts in the field of behavioral health. It does so by translating the articulated constructs into discrete information elements and value sets that can be leveraged in information systems.

2.2 Background

Psychometric instruments, inventories, surveys, and questionnaires are widely accepted tools in the field of behavioral health[28]. They are used in primary psychological research to screen and allocate study participants, and to operationalize study variables. In patient care, they are used to screen for and diagnose mental health conditions. They are used to obtain information about the nature and severity of psychological symptoms, and to select appropriate treatments. They are also used to monitor patients’ response to treatment, and to detect potential adverse reactions to treatment[10]. In clinical research, instruments are used to assess the efficacy and effectiveness of specific therapeutic interventions and programs and to compare the relative effectiveness of two or more interventions. In quality improvement and general oversight, they are used to identify gaps in quality and guide performance improvement programs.

The existence of a robust information model and controlled vocabulary that can be used to enhance terminology products would have a number of advantages. In publication indexing systems, such as PubMed[29] and PsycINFO®[30], the existence of standard concepts for unambiguously identifying psychological instruments would enable the more accurate and efficient indexing of instrument-related publications. In instrument databases, the availability of a controlled vocabulary for key fields would enable more targeted and reliable access to instruments with specific properties. In EHRs, the availability of codes for instruments, instrument components, instrument-related procedures, and instrument results would enable the capture, sharing, and aggregation of instrument-related data. It would also enable the implementation of instrument-related order sets and clinical decision support.

This study is based on the assumption that representations of psychological instruments currently available in the healthcare enterprise accurately reflect the instrument-related information needs of key healthcare stakeholders. Based on this assumption, we believed that identifying, extracting, and harmonizing elements of these models will allow us to generate a robust information model that accurately reflects the current information needs of this domain.

2.2.1 Datasources

Data for the current study came from twelve information products currently used to model, manage, or store information related to psychological assessment instruments. The products evaluated for this study included standard terminologies, ontologies, instrument databases, and thesauri. Standard terminologies used in the current study are CDISC's QS Terminology[21,22], SNOMED CT[24], and LOINC[23]. Two ontologies containing substantial representations of instruments were also included. These are CognitiveAtlas[31,32] and the DOLCE Core Ontology of Instruments[33]. Two controlled vocabularies used in two publication indexing systems, MeSH, and the *Thesaurus of Psychological Index Terms*[34] were also used. These vocabularies contain both implicit instrument-related concepts and a substantial number of terms related to instruments and instrument administration. Finally, four instrument databases that index both instruments and instrument metadata were included. These databases are ETS[35] (Educational Testing Service's TestLink database), HaPI[36] (Health and Psychosocial Instruments), MMY-TIP[37] (Mental Measures Yearbook with Tests in Print), and the newest APA database, PsycTESTS®[38].

2.2.2 Procedure

To identify potential datasources for the study, we performed an initial scan for instrument databases, models, terminologies, ontologies, and encyclopedias. We searched for instrument encyclopedias and databases using a list of resources on the University of Minnesota's library web site[39]. This site contained links to a number of resources for obtaining instrument related information, which we further explored. To identify ontologies and terminologies containing a substantial number of instrument-related concepts, we searched BioPortal[40] and UMLS[41] using the strings 'psych', 'instrument', 'questionnaire', and 'assessment'. To identify publications related to modeling instruments, we searched PubMed[29] and PsycINFO®[30] for publications indexed using any combination of the terms 'terminology', 'ontology', 'vocabulary', and 'nomenclature'; and any combination of the terms 'assessment', 'psychological instrument' and 'psychometrics'.

Using this method, we identified nineteen candidate sources. All non-electronic and microfiche-based sources were excluded, leaving fifteen datasources. Of these datasources, two were duplicate sources of a third, and one contained only a handful of high-level concepts related to instruments, and were thus excluded. This resulted in a final set of twelve datasources. These sources are depicted in Table 2.1.

Table 2.1: Datasources for Instrument Model

Data Source	Maintenance Organization
CDISC Questionnaire (QS) Terminology	National Cancer Institute Enterprise Vocabulary Services (NCI-EVS)
CognitiveAtlas	UCLA Consortium for Neuropsychiatric Phenomics
DOLCE Core Ontology of Instruments	DOLCE
ETS Test Collection	Educational Testing Services
Health and Psychosocial Instruments (HaPI)	Behavioral Measurement Database Services
LOINC®	Regenstrief Institute
MeSH	National Library of Medicine
MMY-TIP	Buros Institute for Mental Measurements
PsycINFO®	American Psychological Association
PsycTESTS®	American Psychological Association
SNOMED CT®	International Health Terminology Standards Development Organization
Thesaurus of Psychological Index Terms	American Psychological Association

CDISC= Clinical Data Interchange Standards Consortium; ETS=Educational Testing Services; LOINC=Logical Observations, Identifiers, Names, and Codes; MeSH=Medical Subject Headings; MMY-TIP=Mental Measurements Yearbook with Tests in Print; SNOMED CT=Systematized Nomenclature of Medicine – Clinical Terms

We manually reviewed each datasource. For each source, we identified instrument-related information elements and entered them into a spreadsheet. The procedure used to identify information elements varied by datasource type. For terminologies, ontologies, and controlled vocabularies, we identified elements based on explicitly defined concepts. We also inferred information elements based on relationships between concepts. For example, in MeSH, the concept ‘psychologic test’ had child concepts ‘language test’, ‘aptitude test’ and ‘personality test’. From this relationship we identified the inferred concept ‘instrument content domain’. To identify elements in instrument databases, we examined associated field guides and technical documents, as well as the databases themselves. All databases examined had at least one generic narrative field containing multiple information elements (e.g., “notes”, “comments”, or “summary” fields). For these fields, we identified and abstracted each of the individual information elements contained in the field.

Once all information element lists were complete, we harmonized them to create a master list of elements. Using the master list, we created a matrix of information element by datasource (Appendix A). We then coded each datasource according to whether it contained an *explicit*, *implicit*, or *no representation* of each of the identified elements.

For terminologies, ontologies, and controlled vocabularies, an information element was coded ‘explicit’ if the terminology contained either a concept or term corresponding to the element in question. It was coded ‘implicit’ if the terminology did not contain a corresponding concept or term, but the information element could be inferred based on a relationship between concepts (e.g., the previously described MeSH element 'instrument content domain'). For encyclopedias and databases, an information element was coded 'explicit' if the datasource contained a discrete field for the element in question. It was coded 'implicit' if the element was included in the datasource, but did not have a discrete, dedicated field (e.g., element 'instrument item count' included in a generic 'summary' field). For all datasources, an information element was coded 'no representation' if the datasource contained no concept, term, or field corresponding to the element in question.

Table 2.2: Instrument Coding

Category	Terminologies	Databases
Explicit	Terminology contains a discrete concept for the target attribute.	The database contains a discrete field for the target attribute.
Implicit	There is no specific concept for the attribute in question, but the concept can be inferred from relationships among concepts, or concept qualifiers.	The database does not contain a discrete field for the target attribute, but the attribute is included in one of the database fields.
No Representation	The attribute is not currently represented anywhere in the terminology.	The attribute is not captured in any field in the database.

For one set of datasources – instrument databases – we found significant variation in information elements across records. While most database field guides contained an extensive list of available fields, many of these fields appeared in only a subset of the records we retrieved. Moreover, the information elements in narrative fields (e.g. 'notes', 'summary', and 'comments'), were highly variable across records. Given this extensive variability, we decided to evaluate information elements in databases using a sample of records, from which we could derive a summary metric.

To generate this sample, we began by creating a master list of mental health conditions. Using the NIMH list of mental health topics[25] as our primary source, we generated a list of sixteen conditions. To this list, we added four constructs that crossed diagnostic boundaries and which we believed to be particularly relevant to mental health. These are depicted in Table 2.3. For each condition or construct, we selected instruments based on recommendations provided by organizations dedicated to the study or treatment of the target condition or construct. We selected three to five instruments for each condition, which resulted in a master set of 65 instruments (Appendix D). Instruments that were not included in at least three of the four databases (ETS, HaPI, MMY-TIP, and PsycTESTS®) were excluded. This resulted in the final set of instruments (n = 30) depicted in Table 2.4.

We manually reviewed the corresponding record for each instrument in each of the four databases. Using the matrix of instrument elements by database, we coded each record according to the scheme depicted in Table 2.2. A total 102 records were coded. Sixteen records (15%), including four from each database, were randomly selected and coded by a second rater. Interrater reliability was calculated for this sample, and found to be adequate ($\kappa = 0.88$).

Table 2.3: Instrument Sample Details

Construct	Type	Source
Generalized Anxiety Disorder	Condition	NIMH
Obsessive-Compulsive Disorder (OCD)	Condition	NIMH
Panic Disorder	Condition	NIMH
Post-Traumatic Stress Disorder (PTSD)	Condition	NIMH
Social Phobia (Social Anxiety Disorder)	Condition	NIMH
Generalized Anxiety Disorder	Condition	NIMH
Attention Deficit Hyperactivity Disorder (ADHD, ADD)	Condition	NIMH
Autism Spectrum Disorders (Pervasive Developmental Disorders)	Condition	NIMH
Bipolar Disorder (Manic-Depressive Illness)	Condition	NIMH
Borderline Personality Disorder	Condition	NIMH
Depression	Condition	NIMH
Eating Disorders	Condition	NIMH
HIV & AIDS	Condition	NIMH
Schizophrenia	Condition	NIMH
Substance Use	Condition	NIMH
Suicide Prevention	Condition	NIMH
Personality / Self	Construct	Author Consensus
Impulsivity	Construct	Author Consensus
Emotion	Construct	Author Consensus
Extra Pyramidal Side Effects	Construct	Author Consensus

After all records in the sample were coded, we generated a summary code for each information element in each database. A summary code of ‘explicit’ was assigned when the information element was coded ‘explicit’ for more than half of the records in the sample. An element was coded ‘implicit’ if it was coded ‘explicit’ for half or fewer of the records, but represented (i.e., coded either ‘implicit’ or ‘explicit’) for more than half of the records in the sample. Finally, an element was coded ‘no representation’ only if it was coded ‘no representation’ for all records in the database. To accommodate information elements in databases that were included in fewer than half of the records, we added a fourth code (“inconsistent”). This code was used to identify those information elements that appeared in too many records to be considered ‘not represented’ and too few records to be considered part of the databases’ overall information model.

Table 2.4: Instrument Sample

Instrument	Category
Addiction Severity Index (ASI)	Substance Use
Agoraphobic Cognitions Questionnaire (ACQ)	Anxiety Disorders
Barratt Impulsiveness Scale (BIS-11)	Impulsivity
Beck Anxiety Inventory (BAI)	Anxiety Disorders
Beck Depression Inventory (BDI)	Depression
Beck Scale for Suicidal Ideation (BSS)	Suicide Prevention
Brief Psychiatric Rating Scale (BPRS)	Symptoms / Functioning
Bulimia Test Revised (BULIT-R)	Eating Disorders
Carroll Rating Scale for Depression	Depression
Child-Behavior Checklist (CBCL)	Symptoms / Functioning
Clinician-Administered PTSD Scale (CAPS)	Anxiety Disorders
Davidson Trauma Scale (DTI)	Anxiety Disorders
Dimensional Assessment of Personality Pathology (DAPP)	BPD
Eating Disorder Inventory (EDI)	Eating Disorders
Eating Disorders Examination (EDE)	Eating Disorders
Hamilton Rating Scale for Depression	Depression
Hostility Inventory (BDHI)	BPD
Michigan Assessment Screening Test - Alcohol Drug (MAST-AD)	Substance Use
Mini-Mental State Examination (MMSE)	Symptoms / Functioning
Penn State Worry Questionnaire (PSWQ)	Anxiety Disorders
Positive Affect - Negative Affect Scale (PANAS)	Emotion
Positive and Negative Syndrome Scale (PANSS)	Schizophrenia
Psychopathy Checklist - Revised (PCL-R)	Personality
Diagnostic Interview for Borderline Personality Disorder (DIB-R)	BPD
Rorschach	Personality
Social Phobia and Anxiety Inventory (SPAI)	Anxiety Disorders
State-Trait Anxiety Inventory (STAI)	Anxiety Disorders
Symptom-Checklist-90-Revised (SCL-90-R)	Symptoms / Functioning
Thematic Apperception Test (TAT)	Personality
Yale-Brown Obsessive Compulsive Scale (YBOCS)	Anxiety Disorders

2.3 Results and Discussion

2.3.1 The Harmonized Set of Instrument-Related Information Elements

We identified a total of 86 information elements related to psychological assessment instruments across twelve datasources. We organized these information elements into eight categories, according to the class of information represented by the elements. These categories are depicted in Table 2.5 and Figure 2.1. A detailed list of elements by category is presented in Appendix B.

Table 2.5: Classes of Information Elements in Existing Representations of Assessment Instruments

Information Class	N	Description
Instrument Scientific Structure	26	Elements related to the quality of the instrument relative to its designed objectives, such as healthcare domain, constructs measured and psychometric properties in various samples
Instrument Access and Logistics	15	Elements related to selecting, procuring and using an instrument for a specific purpose, such as publisher information, permissions required, cost, administration time, and qualifications required
Instrument Physical Structure	12	Elements related to the stimuli and organization of stimuli included in an instrument, such as scales, subscales, items, and item response elements
Instrument Identifiers	9	Elements related to uniquely identifying a specific instrument instance, such as the instrument name, revision number, language translation, publication year and author
Instrument Target Population	8	Elements related to the target population of the instrument, such as gender, culture, social roles, exposure history and healthcare status
Instrument Administration	6	Elements related to instrument administration, such as instrument media, administration contexts, item delivery method and administration protocols
Instrument Scoring and Interpretation	6	Elements related to scoring and interpreting results of the instrument, such as scores yielded, score interpretation algorithms and available scoring norms
Instrument Related Publications	4	Elements related to publications in which the instrument was evaluated, used, or described such as the primary source, second source, and publication sample

We performed several analyses of these data. First, we analyzed the general distribution of elements across information classes. Not surprisingly, the single largest class of elements found across healthcare representations is that related to the scientific structure of instruments. Almost one third (30%) of all elements identified fall into this class (Figure 2.1). The next two largest classes of elements are those related to instrument access and logistics (17%) and those related to the physical structure (12%) of instruments. Together, these three classes account for nearly two thirds (62%) of the elements identified.

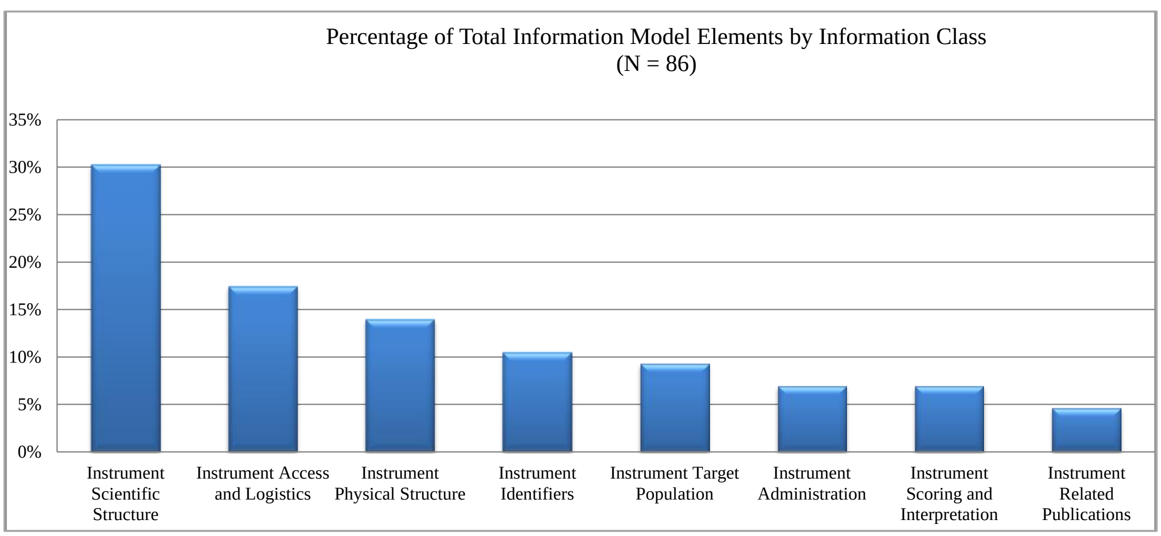

Figure 2.1: Number of Instrument Elements by Information Class

Next, we analyzed the general distribution of elements across datasources. The total number of information elements per datasource ranges from 4 (5%) elements in MeSH, to 60 (70%) elements in PsycINFO®. The percentage of total information model elements by datasource is depicted in Figure 2.2. Not surprisingly, datasources containing the largest number of information elements are instrument databases, and those containing the smallest number are controlled vocabularies used for publication indexing.

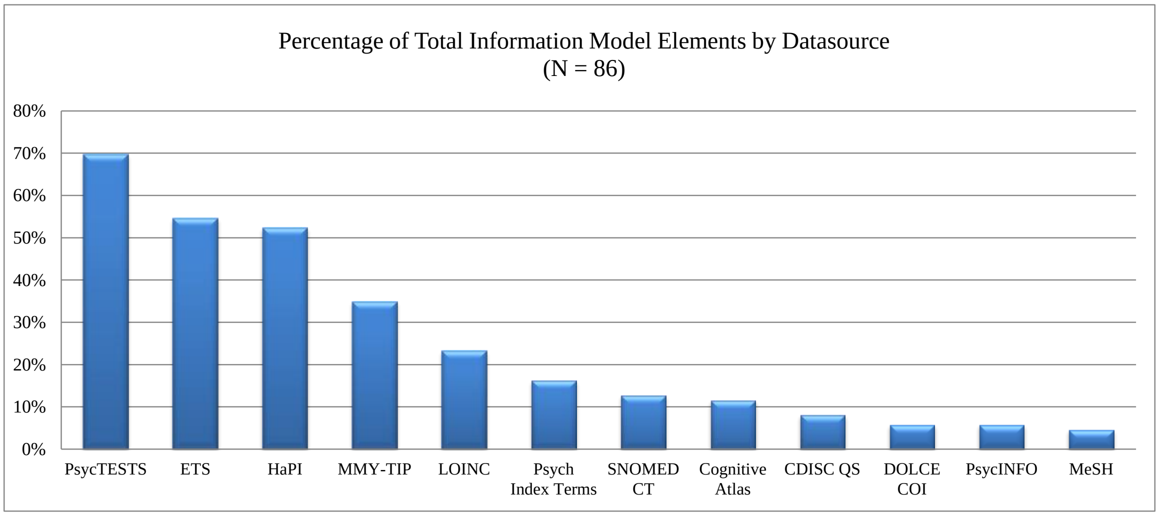

Figure 2.2: Percentage of Information Model Elements Included in Each Datasource

The total number of information elements represented in a given datasource provides a general, high-level view of the variety of instrument-related information contained in each datasource. It also offers insight into how extensively each datasource covers the overall domain of instruments. However, it provides little insight into the quality or utility of the datasource vis-à-vis its specific role in the healthcare enterprise. In order to assess the quality of representations, we performed several additional analyses.

2.3.2 The Representational Characteristics of Instrument-Related Information Elements

A summary view of the representational characteristics of information elements identified in this study is presented in Table 2.6. Using the information class as the unit of analysis, we calculated the number and percentage of information elements represented in each form for each information class. Not surprisingly, instrument identifiers were the class of elements most likely to be explicitly represented and least likely to not be represented. Along with instrument identifiers, information elements related to instrument scoring and interpretation, instrument-related publications, and instrument physical structure were the classes most likely to be explicitly represented.

Table 2.6 : Frequency of Model Element Representational Form Found in Instrument Representations

Information Category	Consistently Explicit		Consistently Implicit		Inconsistently Represented		No Representation	
	N	%	N	%	N	%	N	%
Instrument Identifiers	25	23%	11	10%	12	11%	60	56%
Instrument Scoring and Interpretation	11	15%	2	3%	15	21%	44	61%
Instrument Related Publications	6	13%	2	4%	2	4%	38	79%
Instrument Physical Structure	17	12%	10	7%	16	11%	101	70%
Instrument Scientific Structure	10	3%	22	7%	42	13%	238	76%
Instrument Administration	2	3%	3	4%	11	15%	56	78%
Instrument Access and Logistics	6	3%	4	2%	11	6%	159	88%
Instrument Target Population	1	1%	3	3%	14	15%	78	81%
	78	8%	57	5%	123	12%	774	75%

Information elements related to *instrument administration*, *instrument access and logistics*, and in*strument target populations* are the classes least likely to be explicitly represented, and most likely to be represented inconsistently. The most troubling finding of this analysis, however, is that fewer than 3% of all elements in the class most critical to assessing the in-vivo performance of an instrument (i.e., *instrument scientific structure*) are explicitly represented. Consequently, any user attempting to access information related to these four classes of information will not likely be able to easily or reliably retrieve the target information.

Using the individual datasource as the unit of analysis, we performed two additional analyses. First, we analyzed the distribution of information elements by representational form (i.e., ‘implicit’, ‘explicit’, and ‘inconsistent’). The results are depicted in Figure 2.3. From this graph, it is clear that while instrument databases contain the largest number of information elements, the majority of these elements are inconsistently represented. This finding is primarily attributable to the frequent use of unstructured, narrative fields in instrument databases. These fields typically contain a large number of information elements. However, these elements vary extensively across records.

Even those fields having a fairly specific scope of information contain a variable and inconsistently represented set of information elements. For example, PsycTESTS® contains two discrete fields for capturing psychometric properties of instruments: *validity*, and *reliability*. In our sample, the *validity* field included information elements corresponding to six distinct types of validity, and the *reliability* field included information elements corresponding to four types of reliability. While at least one information element was included in most records for both *validity* (81%) and *reliability* (88%), on average, each of the specific validity elements was reported for fewer than one fifth (19%) of all records, and each of the reliability elements was reported for just over one third (35%) of all records. Consequently, of the twelve psychometric property information elements, only two (*validity* and *reliability*) were classified as ‘explicit’. One information element (*internal consistency*) was classified as ‘implicit’ and the remaining nine elements were classified as ‘inconsistently represented’.

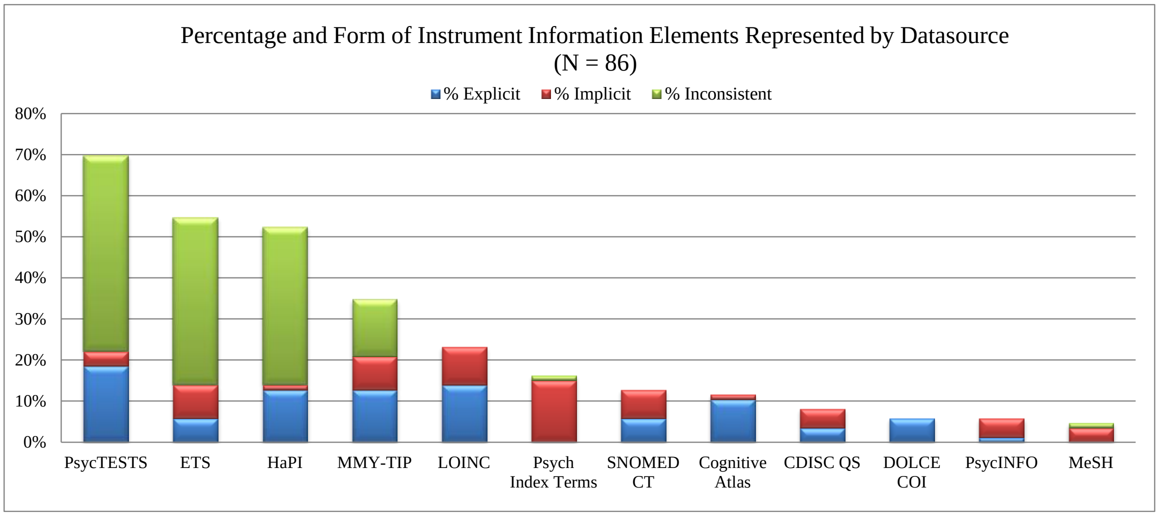

Figure 2.3: Percentage and Form of Instrument Information Elements by Datasource

For the second analysis, we excluded all information elements that were not consistently represented. This view of the data provides a clearer picture of the overall quality of each datasource vis-à-vis its ability to represent the elements in its domain of coverage. As seen in Figure 2.4, the information product containing the single largest number of consistently represented elements is LOINC. This is not surprising, given LOINC's status as the nationally designated terminology for representing psychological assessments.

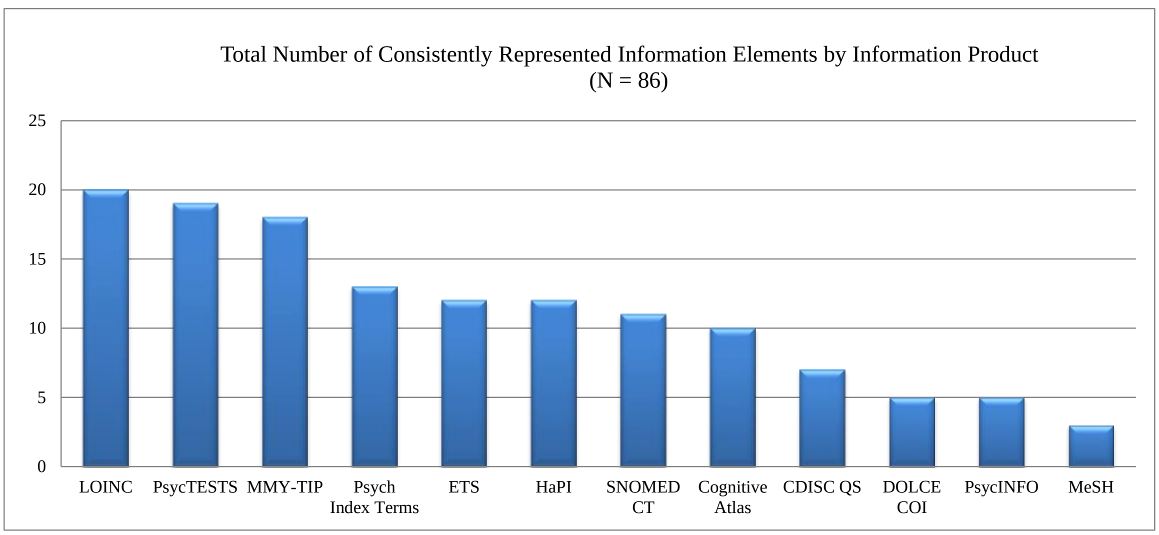

Figure 2.4: Total Number of Implicit and Explicit Information Elements by Datasource

2.3.3 Gaps in Terminology Structure across the Healthcare Enterprise

For the final analysis, we examined potential gaps across the healthcare enterprise vis-à-vis the capacity of existing terminologies to represent specific classes of instrument-related information. First, using the information product as the unit of analysis, we identified those products that covered information elements in a given class more robustly than all other products in the enterprise. Results of these analyses revealed that for three of the eight (38%) information classes (*instrument identifiers*, *instrument administration*, and *instrument populations*) there was no single information product that specialized in representing these elements.

Next, using the information class as the unit of analysis, we identified those classes of information represented by a given information product more robustly than all other products in the enterprise. This analysis revealed that LOINC is the terminology most capable of representing information elements related to *instrument physical structure* and *instrument scoring and interpretation*, while instrument databases are most capable of representing information elements related to *instrument administration*, *instrument access and logics*, *instrument scientific properties*, and *instrument target populations*.

Three important conclusions can be drawn for these analyses. First, while all information products contain some representation of instrument identifiers, no information product evaluated is currently capable of fully representing the *instrument identifier* class of information. Secondly, while several available information products include robust representations of the *instrument administration*, *instrument target populations*, and *instrument scientific properties* classes, all of these products are instrument databases. That is, the information elements included in these products can support the retrieval and indexing of instrument related information, but cannot be used to code information in electronic information systems.

Finally, of the twelve information sources in our sample, only three (LOINC, SNOMED CT, and QS Terminology) can be implemented in clinical information systems (i.e., used to code data). Consequently, elements in all other information products that must be captured in clinical or research activities must first be incorporated into one of these three terminologies. To support this effort, we developed three information models that integrate representations of information elements across all twelve information products evaluated.

2.3.4 Instrument Information Models

Using the information elements identified in our analysis of the twelve information products, we developed three distinct information models. The first model includes information elements required to develop a standard, normalized naming convention for psychological instruments. The second model includes information elements required to represent the physical structure, administration procedures, and target populations of psychological instruments. The third, and final, model includes information elements required to represent the scientific properties and in-vivo performance of instruments. Each information model includes a small number of sample value set items for every information element to which a value can be assigned.

2.3.4.1 The Instrument Identification Model

The instrument identification model includes elements required to develop a standardized naming convention for psychological assessment instruments. This model consists of a number of distinct instrument identifiers (see Figure 2.5). Instrument identifiers are information elements that allow one to unambiguously identify an instrument. These include the *instrument name*, *instrument acronym*, and *alternative names*, as well as the *author*, *publication date* and *primary reference* for the publication in which the instrument was initially described.

Instruments are often locally modified to meet the needs of a specific study or population[42]. When instruments are updated, revised or modified, a number of additional information elements apply to the instrument. These attributes include the *revision identifier, study modification revision,* and *healthcare objective revision.* The *revision identifier* is used to describe the chronological revision number of the instrument (e.g., "Barratt Impulsiveness Scale 11A"). The *study modification revision* is the name of the study for which the instrument was modified. The *healthcare objective revision* describes the function for which this particular version of the instrument is designed (e.g., epidemiological, clinical, research). The elements *number of scales* and *number of items* are also frequently used to identify a specific instance of an instrument, as modifications often produce an instrument with a different physical structure.

Finally, many instruments are developed by modifying and/or combining components of one or more existing instruments to create a new instrument. The element *derived from* is used to accommodate this information. In the current model, we distinguish between *simple modifications* (i.e., changes not intended to change the substantive properties of the instrument) and *adaptations* (i.e., changes that significantly alter the operational definition of the construct or constructs measured by the instrument). Both types of changes have a type (length, language, setting, construct, scale, response options, etc.) and a source instrument (i.e., the instrument to which, or from which, the changes were made). However, under the current scheme, simple modifications are modeled using the *version* element of the *instrument identifiers* class, while adaptions are modeled using the *derived from* element of the *scientific structure* class.

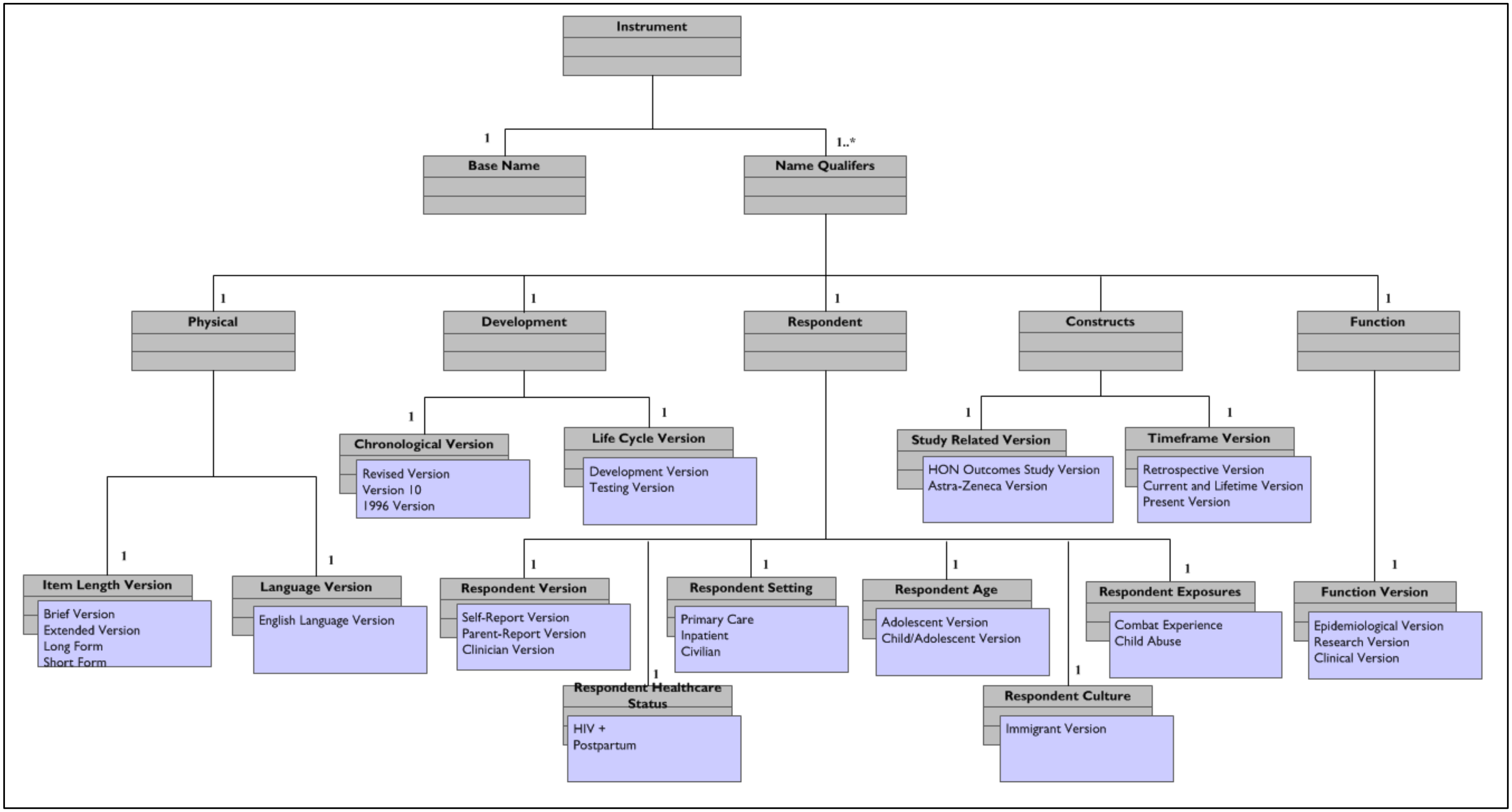

Figure 2.5: Entities Related to the Development of a Standard, Normalized Naming Convention

2.3.4.2 The Integrated Information Model

2.3.4.2.1 Instrument Physical Structure

A number of elements in the model relate to the physical structure of instruments. These include *instrument* s*cale, instrument items* and *item responses*. The *scale name* and *scale acronym* attributes have been added to both the instrument scale entity and instrument subscale entity. While these attribute are less commonly made explicit in the existing information models we reviewed, we opted to add the attribute *short item name, item text,* and *item response text* to the item entity. The *instrument readability level* is an intrinsic property of instruments, and is an important information element for instrument selection.

In the current model, a *scale* is defined as a named or unnamed subset of one or more items within an instrument. Scales have a recursive relationship. That is, a scale can contain additional scales. In this model, the instrument itself is considered the top level scale. A scale is necessarily associated with one or more constructs and one or more scores. The *score* may be a quantitative metric or a qualitative one (e.g., the “dismissive attachment style” of the Adult Attachment Interview). A scale differs from a subscale, in that there is no single subset of items (i.e., scale) other than the instrument itself, to which all scale items belong.

A scale consists of one or more items. An *item* is the most granular element of an instrument upon which a discrete evaluation can be made. In a questionnaire, an item is a single question. In a non-questionnaire type test, such as the *Rorschach Test,* an item is the smallest component of the instrument to which a score is assigned. Each item has a specific item response method. The item response consists of both the *response method* (e.g., Likert Scale, Q-sort, T/F, etc.) and the *response format*.

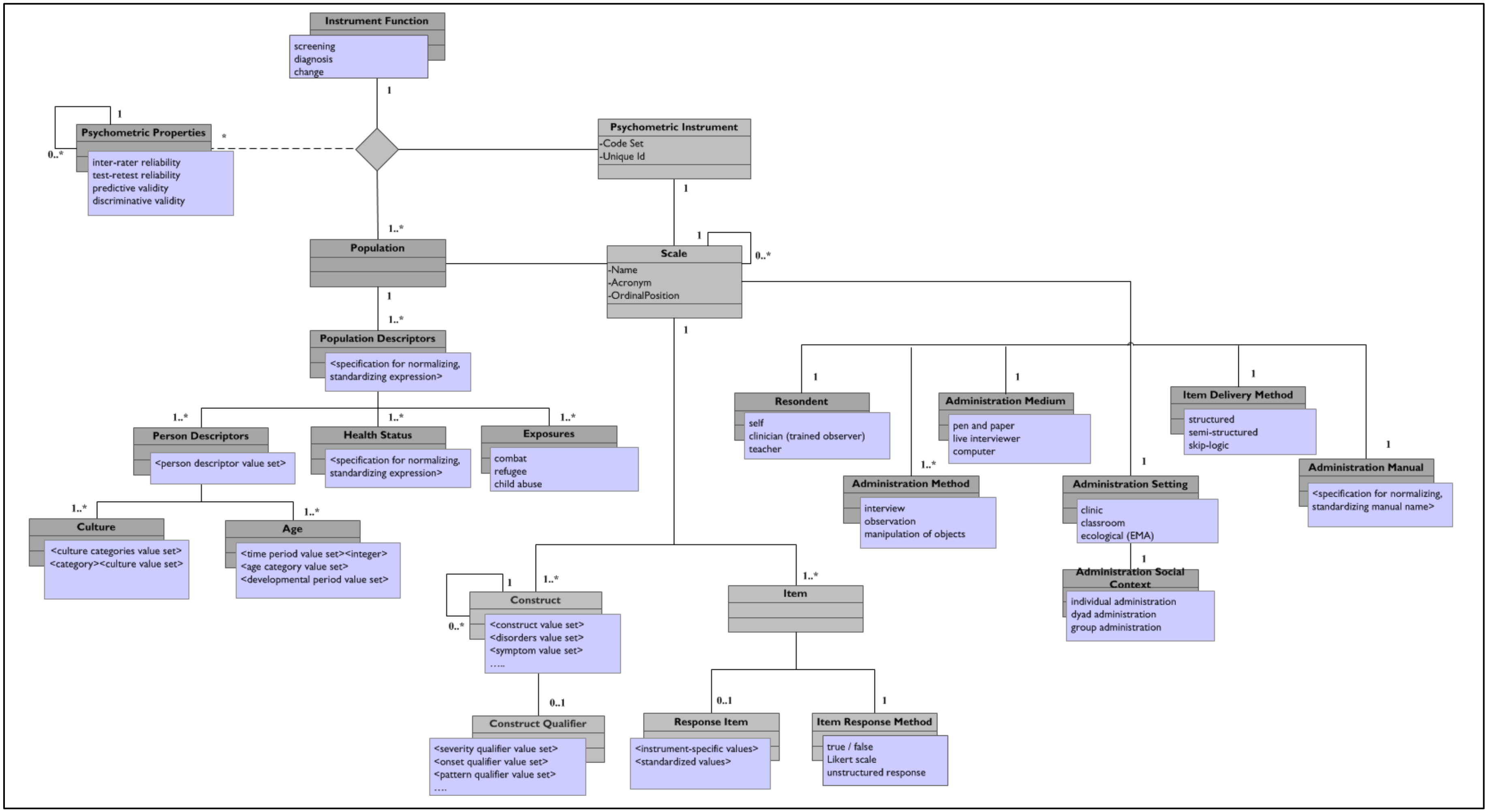

Figure 2.6: Integrated Model of Psychological Assessment Instruments and Instrument-Related Activities

2.3.4.2.2 Instrument Scientific Structure

The scientific structure of an instrument includes elements that are frequently defined during the development and ongoing evaluation of instrument performance as it is used in research and clinical practice. The elements *developmental approach*, *developmental model*, and *construct access method* refer to the theoretical foundation of the test. The *developmental approach* refers to whether the scales and items in the instrument were developed empirically, or based on a theoretical conceptualization of the construct to be measured. The element *developmental model* refines and qualifiers the *developmental approach* element, by describing either the theoretical conceptualization upon which a theoretically-derived instrument is based or describing the empirical method used to develop the instrument.

The *developmental life-cycle* of an instrument refers to the extent to which the scientific properties of an instrument are known. An instrument identified as a 'development', or 'beta' version is an instrument that has been released, but whose psychometric properties are not yet fully known (analogous to a drug that is released in an experimental stage).

The *construct access metho*d refers to the theoretical approach used to obtain information about the construct being measured. The traditional, high-level classification of tests as either 'objective' or 'projective'[43,44] are examples of the two most commonly used construct access method values. Other candidate values include those defined by Bornstein[44] in his process-based framework for classifying psychological tests. These include 'self-attribution', 'stimulus attribution', 'performance-based', 'constructive', 'observational' and 'informant-report'. More specific candidate values might include items such as those included in the MeSH terminology, such as 'semantic differential tests' or 'inkblot tests'.

The *construct target time-frame* refers to the time period for which the construct in question is being measured. Value set items include both explicit time periods such as 'today' and 'during the prior 2 weeks', and socially-defined periods such as 'during childhood', 'during pregnancy', and 'during the 6 weeks after the loved one died'. The distinction between these types of time-frames is important, because terminologies often lack concepts for representing psychologically and socially-relevant time periods (see, for example, discussions in early evaluations of LOINC[45]).

Finally, the *respondent*, and his or her relationship to the subject of the test, is an important information element not only for describing a specific instance of an instrument (i.e., 'self-report version', 'parent version'), but also for interpreting results[43].

2.3.4.2.2.1 Psychometric Properties

Each instrument has associated psychometric properties. These properties, however, are not intrinsic attributes of the instrument, but rather are a function of both the sample and objective for which the instrument is used[46]. The evaluation sample describes the sample used to calculate psychometric property metrics. Each instrument may have one or more evaluation samples, and the metrics reported may be based on pooled data, or individual sample data. As depicted in Figure 2.7, psychometric properties fall into two general categories of ‘validity’ and ‘reliability’. Reliability is generally defined as a measure of the consistency of results obtained between raters, over time, or between instrument components. Validity is generally defined as a measure of the extent to which the instrument measures the construct it purports to measure[47]. Like the more general class *psychometric properties*, neither validity nor reliability can be directly measured, and therefore have no associated metrics or formal algorithms for assigning associated values. Instead, all three of these constructs represent generic classes to which summary impressions can be assigned (e.g., ‘good’, ‘poor’, ‘adequate’). While some general guidelines exist for assigning values to these classes (i.e., inter-rater reliability >= 0.70 is adequate), we are not aware of any formal algorithms for assigning values to these information element classes.

There are many specific types of reliability and validity, however, for which metrics can be calculated. Three types of reliability and six types of validity were identified in our sample. These include ‘test-retest reliability’, ‘inter-rater reliability’ and ‘parallel forms reliability’; as well as ‘construct validity’, ‘content validity’, ‘divergent validity’, ‘convergent validity’, ‘predictive validity’, and ‘profile validity’. All types of reliability and four of the six types of validity can be operationalized and measured. Content validity and construct validity, like the more general class ‘validity’, cannot be directly measured. Each instance of a validity or reliability metric consists of the operationalization of the metric, and a numeric result.

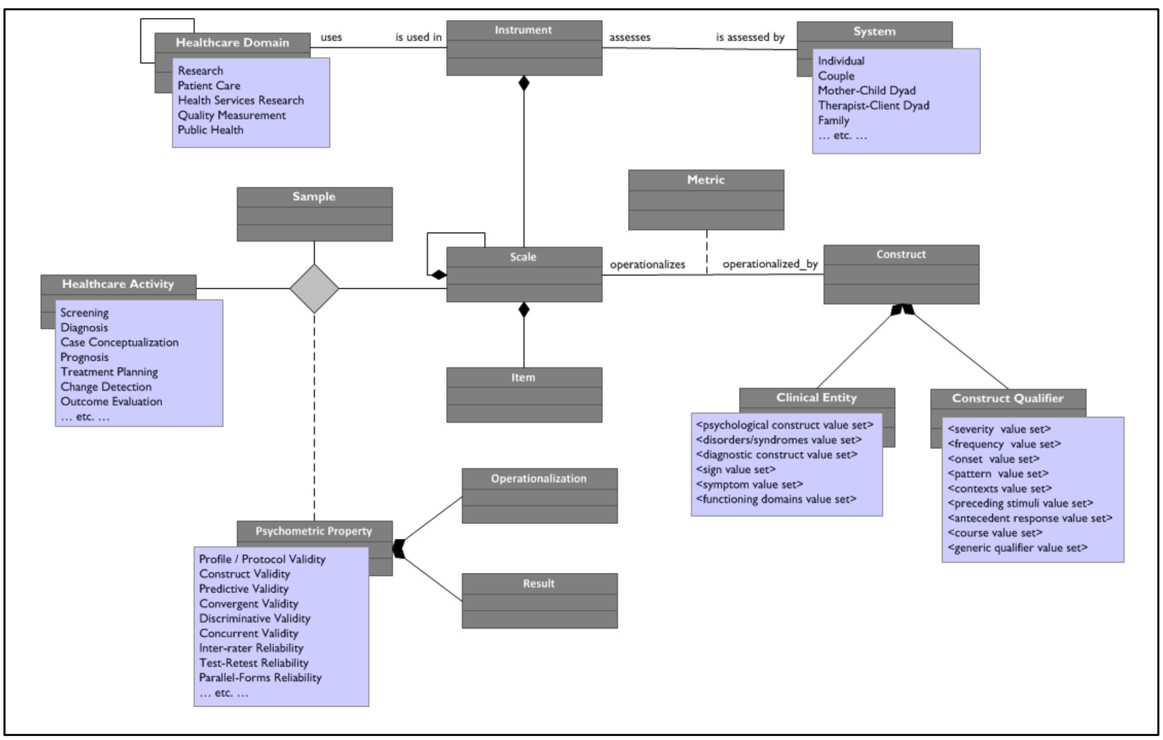

Figure 2.7: Entities Related to the In-Vivo Performance of Psychological Assessment Instruments

2.3.4.2.2.2 Instrument Function

The function of an instrument is defined in terms of the healthcare domain and clinical or research objective(s) for which the instrument was developed, as well as the specific activities that are accomplished using the instrument. It is also defined in terms of the healthcare construct, or constructs, that the instrument was designed to measure.

The *healthcare domain* refers to the specific domain or discipline for which the instrument is developed, or in which the instrument is used. These domains include psychometrics research, primary research, clinical research, healthcare delivery, health services research, epidemiological research, and quality measurement and reporting. The *healthcare objective* refers to the high-level healthcare function, or purpose, for which the instrument is developed or used. Objectives include determining prevalence and incidence of a specific symptom, disorder or syndrome; obtaining a patient history; making a diagnosis; and formulating a case. Other examples of objectives are determining a patients' prognosis, developing a treatment plan, monitoring response to treatment, and evaluating outcomes.

A *healthcare activity* is any activity performed that involves the use of a psychological instrument. These activities include calculating the divergent or convergent validity of an instrument in a specific population; determining the prevalence or incidence of a specific sign, symptom or disorder in a community; and assessing the existence, severity, duration or frequency of a specific symptom. Other examples of instrument-related healthcare activities include monitoring signs or symptoms, and obtaining a standard patient history.

A *behavioral health construct* is the class of behavioral or psychological entities measured by the instrument. A *behavioral health construct* may be an entire psychological "system" (such as personality, emotion, or cognition), a specific psychological construct, a sign or symptom, a disorder or syndrome, a level or area of functioning, or a diagnostic construct. A *construct qualifier* is used to further refine the *behavioral health construct*. Construct qualifiers include constructs related to the presence of symptoms or disorders (e.g., predisposition towards, differentiation/discrimination between); the specific attributes of symptoms or disorders (e.g., severity, frequency, pattern, and change); and specific manifestations (e.g., contexts of occurrence, eliciting stimuli, aggravating factors, and ameliorating factors) or consequences (e.g., impact of functioning) of a symptom, syndrome, or intervention.

2.3.4.2.3 Instrument Target Population

The instrument target population describes the specific population for which the instrument was developed. It represents the population to which the psychometric properties reported in initial evaluation samples are assumed to apply. It is the population for which the specific *developmental model*, *developmental approach* and construct operationalization (through scales, items, item-responses, and administration method) are selected to ensure maximal construct validity. The instrument target population is modeled in terms of specific attributes of the subject of the test (commonly referred to as the respondent). These include intrinsic attributes, such as the *respondent age* and *respondent sex.* They also include socially constructed attributes such as *respondent gender*, *respondent race, respondent culture* and *respondent social role* (e.g., parent, widow, caregiver, refugee, and college student). Other important attributes include the *respondent setting* (e.g., prison, inpatient unit, home, and school); respondent *healthcare status* (e.g., HIV+, pregnant, and psychiatric patients); and *respondent exposures (*e.g., rape, witness to interpersonal violence, and combat experience).

2.3.4.2.4 Instrument Access and Logistics

A number of elements in the model are used to represent and store data related to the logistics of instrument use. Instrument *publication status* describes whether the instrument was ever published and whether it is currently in print and/or available. The *instrument access source, publisher name, publisher address*, and *instrument web site* elements all model information about how the instrument can be accessed. The *administration manual* and *associated materials* elements model additional material that can be accessed to facilitate administration of the instrument. The instrument *price*, *languages available*, *administration time*, *administration accommodations,* and *qualifications* and *permission required* to use the test are all elements that support test selection decisions.

2.3.4.2.5 Instrument Scoring and Interpretation

As described previously, the *respondent*, and his or her relationship to the subject of the test, is an important information element not only for describing a specific instance of an instrument (i.e., self-report version, parent version), but also for interpreting instrument results[43]. The *scores yielded* element, and the *type of score* (e.g., raw score, total score, profile, aggregate) element are also included in this class. In addition to scores, some tests yield *factors* that are distinct from specific scale or subscale scores. These factors may or may not have been defined as part of the original description of the instrument. Finally, standardized tests, or tests that have been normed against some relevant population, allow a respondent's score to be reported in terms of these norms. Publications relating to the derivation of *norms* are also important information elements of this class.

2.3.4.2.6 Instrument Administration

There are a number of entities and attributes in the model related to instrument administration. The *administration method* is defined as the overall approach taken to obtain the measurement data. Potential value-set items for this element include 'interview', 'observation', 'manipulation of objects', and 'interaction with media, environment or technology'. The *administration manual* is a specific, written protocol used to describe a standardized administration process. *Administration accommodations* refer to specific modifications made to the procedure for specific respondent populations. The *administration medium* is the medium in which the instrument or scale is delivered. Potential value-set items for this element include ‘pen-and-paper’, ‘face-to-face live interview’, ‘audio’, etc. The *administration setting* is defined as the setting in which the instrument is administered. Potential value-set items include ‘home’, ‘classroom’, ‘workplace’, ‘work environment’, ‘clinic’, and ‘hospital’. The *administration social context* is the social context in which the instrument is delivered. Potential value-set items include ‘individual’, ‘couple’, and ‘group’.

The *item delivery method* is defined as the method by which individual items are delivered. Potential value-set items for this element include ‘structured’, ‘semi-structured’, ‘skip-logic’ or ‘unstructured’. The *item response method* defines the specific manner in which the respondent is asked to provide his or her response. Potential value-set items include ‘true/false’, ‘Likert scale’, ‘semantic differential’, ‘sentence completion’, and ‘visual analog scale’.

2.3.4.2.7 Instrument Evaluation

Instrument evaluation occurs both formally and informally. Formal instrument evaluation occurs when an instrument is first released. The evaluation is typically reported in a publication, which is generally referred to as the primary source publication. Formal evaluations also occur when a researcher reviews publications using the instrument and synthesizes results from multiple secondary sources. These evaluations may be published in a peer-reviewed journal, or written as an invited review that appears in an instrument review handbook or database.

2.4 Conclusions

In this study, we analyzed representations in twelve information products related to psychological assessment instruments. We identified a set of 86 information elements, across eight classes of information. Our analysis of information sources revealed a great deal of variation in the breadth and depth of information elements covered, as well as the form and consistency of representations. Much of this variation is appropriate, expected, and desirable. That is, the existence of several high-quality terminologies or information sources, each one representing a specific domain of information using explicit information elements with predefined value sets is highly desirable.

Gaps and redundancies, on the other hand, are not desirable. The current analysis revealed both when viewing information products from the perspective of the overall healthcare enterprise. The existence of four instrument databases, all including the same general information elements, and all covered with the same level of consistency and quality represents waste. The resources currently being invested in maintaining four distinct systems could be leveraged to produce a single high quality database, including more explicit coverage of one of the key gaps in current information sources: representations of the scientific properties of instruments.

We identified several gaps in terminologies. Most significantly, none of the terminologies examined is capable of fully representing instrument identifiers. Consequently, there is currently no standard way to unambiguously identify specific psychological assessment instruments in the healthcare enterprise. Moreover, no terminology examined is capable of representing information related to the administration or scientific properties of these instruments. To enable the capture, aggregation, and sharing of these data, standard terminologies must be enhanced to include the classes of information currently included in the ontologies, instrument databases, and controlled vocabularies evaluated in this study.

Finally, more research is needed to address important questions regarding which products are most appropriate for representing which instrument-related information elements. In particular, the current national strategy for representing psychological instruments only in LOINC, and not SNOMED[48], may be worth revisiting. These two terminologies have different representational capabilities, and there are many information elements related to psychological instruments that cannot feasibly be represented in LOINC. LOINC is clearly the ideal terminology for representing information elements related to the physical structure of instruments as well as instrument scoring. However, LOINC is not capable of representing the important class of information related to instrument administration. SNOMED CT, on the other hand, is ideally suited for this purpose. Finally, the critical issue of which terminology is most appropriate for uniquely identifying instruments, and modeling the necessary instrument identification elements must be addressed.

CHAPTER 3

THE ACCURACY AND COMPLETENESS OF DATA IN THE "TESTS AND MEASURES" FIELD OF PSYCINFO®: IMPLICATIONS FOR CLINICAL RESEARCH AND PATIENT CARE

Psychometric instruments and psychological assessments, inventories, and questionnaires are a primary source of objective, standardized data in behavioral health. They provide information that allows providers to move from clinical impression to formal diagnosis and case conceptualization. They guide the selection and delivery of general treatment programs and specific interventions. They are the primary source of objective data upon which all evidence-based practice in this domain is based. Consequently, the ability to efficiently access accurate and complete information about psychological instruments is a high priority to researchers, clinicians, and quality measurement organizations. A number of electronic databases have been developed to catalog instruments and instrument information. In addition, one publication indexing database, PsycINFO®, indexes all instruments used in publications it covers. Through its comprehensive coverage of behavioral health publications and use of a discrete field for enumerating all instruments used in indexed publications, this database offers access to the most robust a set of instrument data currently available. The purpose of this study was to evaluate the quality and completeness of data in this field. Using a sample of 203 publications, we found errors in this field for 56% of the publications evaluated. We discuss the implications of these findings for clinical research and patient care.

3.1 Introduction

Psychometric instruments, inventories, surveys, and questionnaires are widely accepted tools in the field of behavioral health[28]. They are used extensively in primary and clinical research, and less extensively in patient care, quality measurement, and payor oversight. In behavioral and psychological research, they are used to screen and allocate study participants, operationalize study variables, and evaluate outcomes. In clinical research and patient care, they are the preferred method for obtaining standardized measurements of demographic, psychological, functional, and neurological variables.

As a central source of objective data upon which all evidence-based practice in this domain is grounded, the ability to efficiently access accurate and reliable information about psychological instruments is a high priority. Unlike general healthcare, in which standardized instruments (such as blood tests and CT scans) perform relatively consistently across the targets they are designed to measure, psychometric instruments do not have intrinsic scientific properties. Critical attributes of psychometric instruments, such as validity and reliability, can only be defined in relation to the population in which they are used[46,49]. This makes access to instrument performance information a high priority for researchers in the behavioral sciences, and clinicians working in behavioral health.

To meet this need, a number of products have been developed to allow users access to relevant instrument-related information. These products include electronic databases[37,50,38], encyclopedias[51,52], reference manuals[53], and handbooks[54]. Each of these datasources covers a somewhat different range of instruments, and provides access to a somewhat different set of instrument data. The breadth and depth of coverage of these datasources varies. However, each of these sources suffers from a similar shortcoming: the range of coverage is restricted by publisher-imposed criteria.

PsycINFO®[55] is one database to which this shortcoming does not apply. PsycINFO® is the official publication indexing database of the American Psychological Association (APA) and is the most extensive and widely used publication indexing database in mental health and the behavioral sciences[56]. It indexes publications from more than 25,000 journals, including publications from more than 50 countries in 29 languages[57]. In June 2003, PsycINFO® introduced the *Tests and Measures* (TM) field[58], a discrete field created to capture the names of assessment instruments used in empirical studies. According to a newsletter published by the vendor, the field was introduced to allow users to search for all publications in which a particular test or measure was used, regardless of whether the instrument was the topic of the publication. The field was designed to capture the names of those tests "used to conduct the research"[58], and does not capture the names of all tests and measures described in the publication.

Unlike instrument-specific databases that catalog instruments first, and include relevant publications secondarily, PsycINFO® catalogs publications first, and uses the TM field to enumerate all instruments used in the study. There is no filtering process in place – instrument data for each and every publication indexed by PsycINFO® is automatically abstracted and made available. Given the databases' extensive coverage of the domain, PsycINFO® is arguably the most comprehensive and robust source of instrument data currently available. Because of its unique position in the larger set of instrument information products, the quality of data in this field is of immense importance in the overall healthcare enterprise. The ability of researchers, clinicians, and other key healthcare stakeholders to acquire actionable knowledge from this information system, however, is contingent upon the quality of data in this field. The primary aim of the current study is to empirically assess the accuracy, completeness, and accessibility of data in this field. A secondary aim is to describe the overall scope of instrument data captured by the PsycINFO® TM field.

3.2 Methods

3.2.1 Datasources

Data for the current study came from a master study database created by one of the authors (PR) as part of a larger study. The goal of the larger study was to identify a core set of variables relevant to research and treatment for borderline personality disorder (BPD). All data in the study database came from PsycINFO®, and publications indexed by PsycINFO®. The study database contains publication data downloaded from PsycINFO®, including all instrument names included in the TM field. It contains all PsycINFO® metadata (i.e., title, author name, methodology, etc.) for a total of 4,676 publications, and links to a local copy of many of the publications. The metadata file was generated by searching PsycINFO® for all publications indexed using "borderline personality disorder" as either a subject heading or keyword. No other restrictions were placed on the search.

This set of publications clearly represents a convenience sample. However, BPD is a syndrome well known for its extensive breadth and depth of associated symptomology and dysfunction[59]. Therefore, we believe that instruments identified by drawing our sample from this set of publications reflect a sufficiently broad range of instruments, assessing a sufficiently broad range of clinical constructs, to justify use of this sample as the basis of the current evaluation.

3.2.2 Procedure

Publication metadata for all 4,676 publications was downloaded from PsycINFO® and loaded into a Microsoft SQL Server® 2012 study database. A separate instrument database was developed to facilitate the mapping of TM entries to standardized, fully-specified instrument names. Using data from the TM field, a preliminary list of commonly used instruments was generated. This list consisted of core instrument names ("base instruments") that represented a class of instruments (e.g., "Beck Depression Inventory") as opposed to specific instances of instruments (e.g., "Beck Depression Inventory Version II"). Additional tables were designed to capture instrument versions, translations and modifications, as well as instrument scales, subscales, scores and factors. Instruments and instrument components were researched and added to the database as needed (i.e., as they appeared in publications being manually reviewed and coded). Each instrument and instrument component in the database was assigned a unique identifier and a standardized, normalized name.

3.2.2.1 Scope of Instrument Use in Publications Indexed by PsycINFO®

We used two distinct samples from this database in the current study. The first sample (n = 1,936) was used to evaluate the overall scope of data in the PsycINFO® TM field. The sample for this evaluation consisted of all peer reviewed journal articles in our BPD database published between Jan 1, 2004 and Jun 30, 2012. Publications classified by PsycINFO® as "review", "systematic analysis", or "meta-analysis" were excluded. No other exclusions were made. Based on preliminary analysis of data in the TM field, we identified a number of issues that could potentially skew evaluation of the scope of instrument use in publications. One issue was that both instruments and instrument components were abstracted and enumerated in the TM field. In some cases, the instrument component was reported in lieu of the instrument. In other cases, multiple instrument components (e.g., all subscales in a given instrument, or both the instrument and the index corresponding to the overall instrument score) were abstracted. Less frequently, two distinct instruments were reported as a single instrument (e.g., the SCID-I and SCID-II reported as a single entry, "SCID"). Consequently, any analysis that simply summed the TM entries by publication would overestimate the number of instruments used. Therefore, prior to running our analysis, we ran a series of data validation and cleansing routines to adjust for these cases.

3.2.2.2 Accuracy of Data in the PsycINFO® Tests and Measures Field

A second, smaller, sample was used to evaluate the accuracy and completeness of data in the TM field. The sample for this analysis (n = 203) consisted of a set of studies that had been manually processed as part of the larger study. Like the first sample, this sample was limited to peer-reviewed journal articles. Publications classified by PsycINFO® as “review”, “systematic analysis”, or “meta-analysis” were excluded. Because each of the studies was manually reviewed, this set of publications was further limited to those publications available in English. All publications in this sample addressed clinical outcomes for psychosocial interventions for BPD (n = 61), biological correlates of BPD (n = 147), or both (n = 5). Publications in the sample related to psychosocial interventions were published between Jan 1, 2007 and June 30, 2012. Publications related to the biological correlates of BPD were published between Jan 1, 2004 and June 30, 2012.

After an extensive review of data in the TM field, we developed a set of six codes for classifying the accuracy and quality of TM entries. In this study, we report only findings related to errors: false negative errors and false positive errors. To facilitate coding, a Microsoft Access® application was built and used as an interface for filtering, viewing and coding publications (Figure 3.1). For each publication, all instruments used in the study were manually abstracted. Any instrument used in the study, but not included in the TM field, was manually added via the instrument coding subform (depicted in the lower part of Figure 3.1). Each TM entry was classified using one of three codes: false positive, false negative, accurate abstraction. Using information provided by study authors in the methodology and reference sections of the paper, each instrument was mapped to a unique instrument identifier. Instrument identifiers came from the instrument database developed specifically for the current study. Each instrument in the database was assigned a standard, normalized name. The instrument identifier table was linked to the study database so that all mapping could be done from within the study database (Figure 3.2).

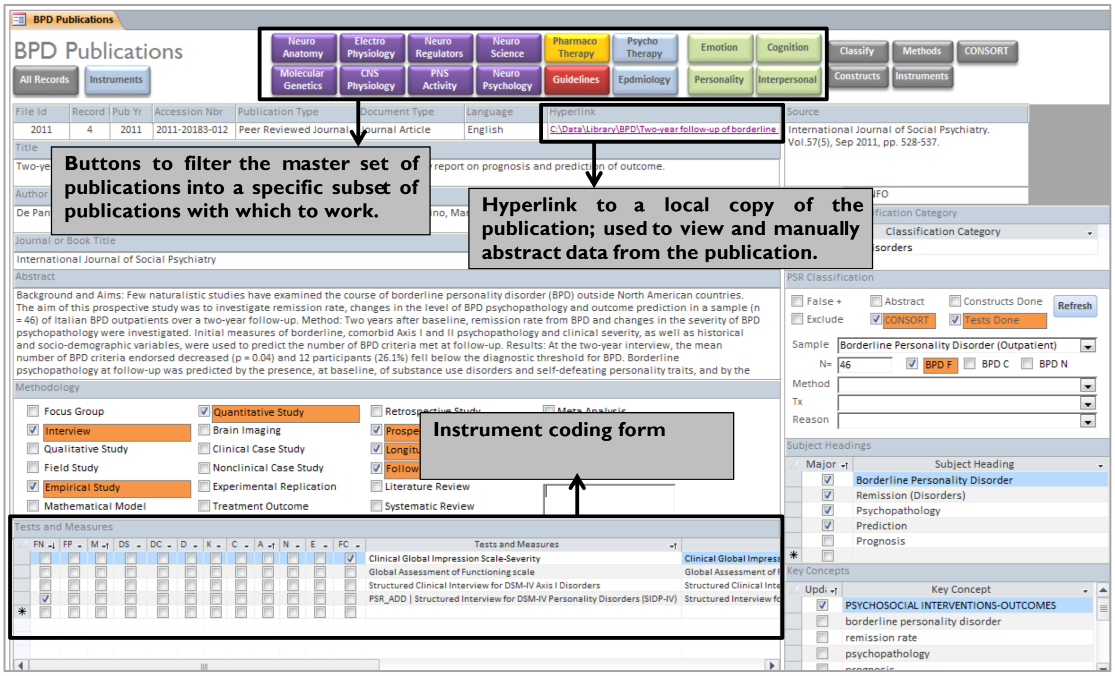

Figure 3.1: Publication Selection, Viewing and Coding Interface

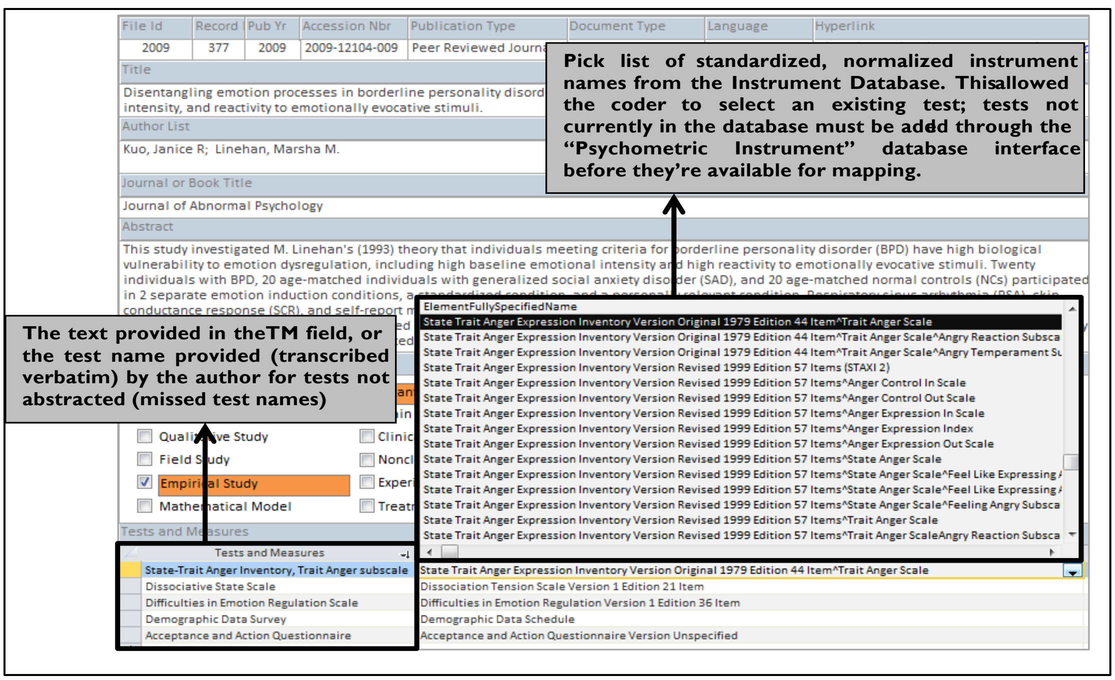

Figure 3.2: Instrument Coding Interface

Twenty-two publications, representing 10% of the total sample, and 175 instrument entries, were randomly selected and coded by a second rater (JBF). A kappa statistic was calculated for classification of abstraction errors across three categories and found to be adequate (weighted κ = 0.80). Virtually all discrepancies between raters occurred in identifying false negatives. All discrepancies between raters were reviewed a second time. The second review resulted in a consensus that all instruments identified as a false negative by one of the raters, but not the other, were in fact false negatives. Interrater agreement was excellent for the identification of false positives ($\kappa = 1.0$). This suggests that the false negatives rate reported in these results may underestimate, but likely will not overestimate, the true false negative rate.

3.3 Results

3.3.1 Scope of Instrument Use in Publications Indexed by PsycINFO®

Our general analysis of the PsycINFO® TM field revealed that two-thirds (66%) of all peer-reviewed publications in our sample (n = 1,936) had at least one entry in the TM field. To better understand the scope of data in the TM field, we explored two trends. The first was the average number of publications using instruments over time. We found a consistent increase in the proportion of publications using formal assessment instruments, from less than half (49%) of all studies published in 2004 (n = 189), to more than three quarters (78%) of those published in the first half of 2012 (n = 88). These results are depicted in Figure 3.3.

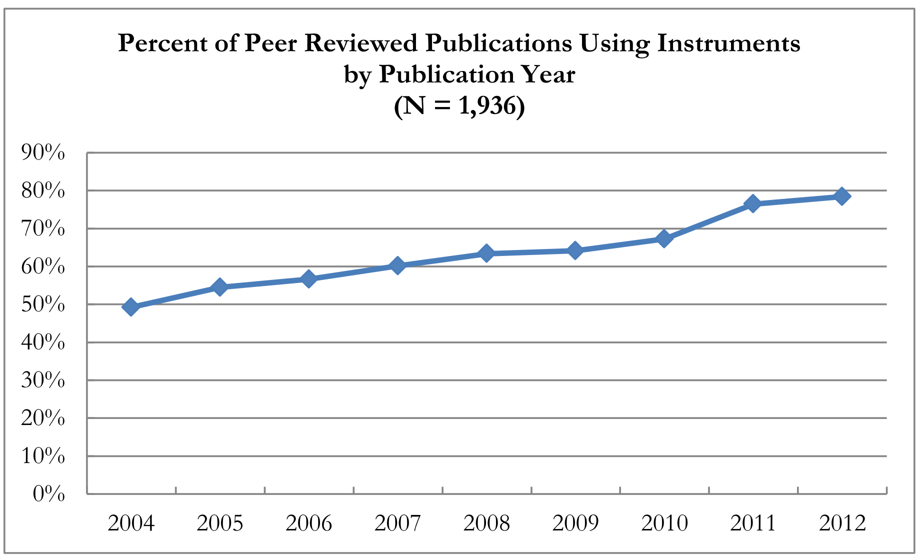

Figure 3.3: Trend in Percent of Peer Reviewed Publications Using Instruments

A second trend in the data was an increase in the average number of instruments used in empirical studies over time. Because the number of publications using instruments increased over time, an overall average would present an artificially inflated picture of instrument use. Therefore, we limited this analysis to only those publications in which at least one instruments was used (i.e., empirical studies). Of 1,275 empirical studies in our sample, the average of number of instruments used increased from 3.7 in 2004 (n = 93) to 5.0 in the first half of 2012 (n = 69).

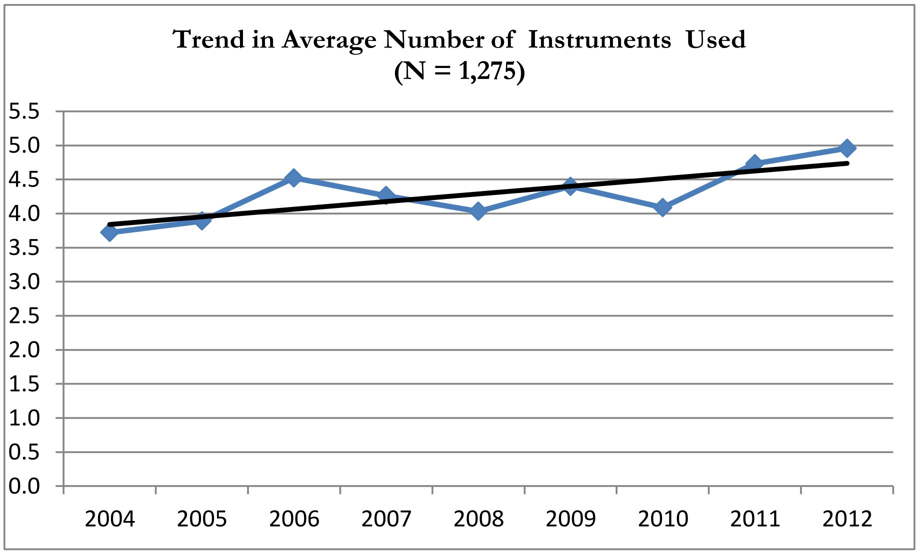

Figure 3.4: Trend in Average Number of Instruments Used in Empirical Studies

3.3.2 Data Accuracy

Our second, smaller sample of 203 publications referenced a total of 1,134 instruments. Of these publications, a total of 116 (57%) contained abstraction errors. The distribution of errors is depicted in Figure 3.5. A total of 109 (54%) publications contained one or more false negative (FN) errors, 15 (7%) contained one or more false positive (FP) errors, and 8 (4%) contained at least one of each type of error. There was an average of 1.2 FN, and 0.1 FP per publication. The false positive errors in this sample were primarily due to abstractors incorrectly abstracting instruments that were described by authors, but not actually used in the study. Less commonly, physiological measures (e.g., "Cardiac Sympathetic Index"[60]) and statistical procedures ("Reliable Change Index"[61]) were incorrectly abstracted as psychological.

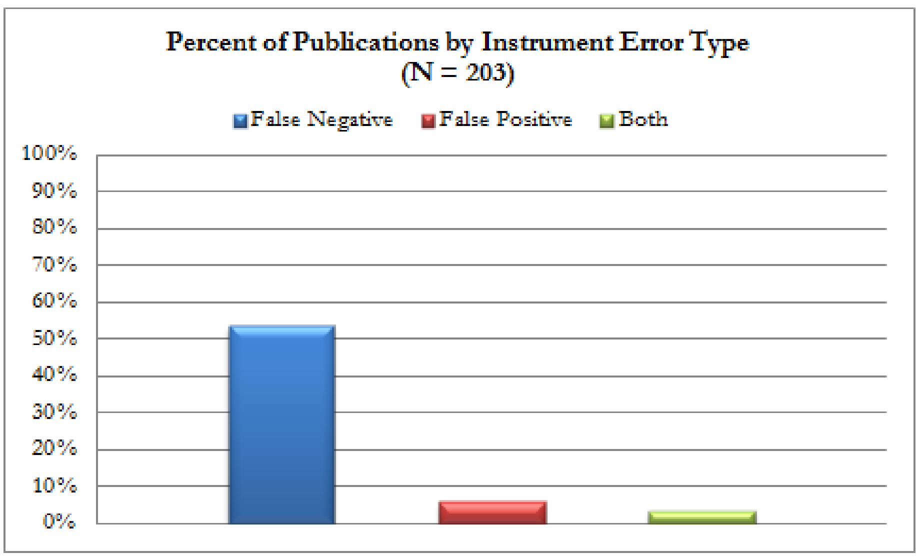

Figure 3.5: Instrument Error Rate in Publications

The distribution of false negatives in our sample is depicted in Figure 3.6. The number of false negatives ranged from 0 (n = 94, 46%) to 8 (n = 1, 0.5%) Of those publication entries containing false negatives, more than half (54%) were missing more than one instrument in the TM field, and 15% were missing 3 or more. The average number of missing instruments in those publication entries containing false negatives was 2.2.

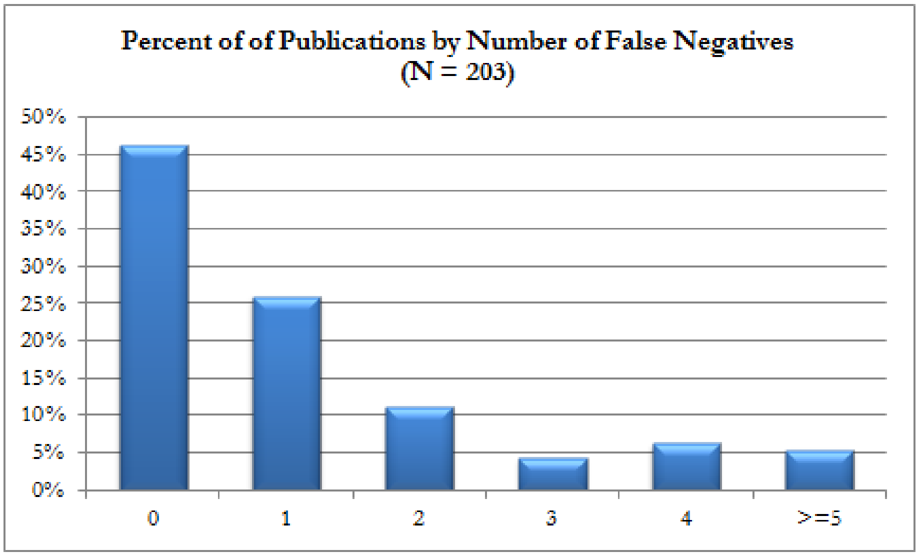

Figure 3.6: False Negative Error Rate

While the error rate per publication is an important metric from the perspective of identifying problems with the abstraction process itself, it is less relevant from the perspective of predicting the quality of information a clinician or researcher will obtain when using data in this field. From this perspective, the important metric is the overall rate of false negative and false positive errors in the entire sample of instruments.

To determine this rate, we calculated the abstraction error rate relative to the total set of tests and measures included in our sample. Of 1,134 instruments evaluated, 877 (78%) of the entries were correctly abstracted, 240 (21%) were missed (i.e., false negatives), and 17 (1.5%) were incorrectly identified as having been used in the study (i.e., false positives). These results are depicted in Figure 3.7.

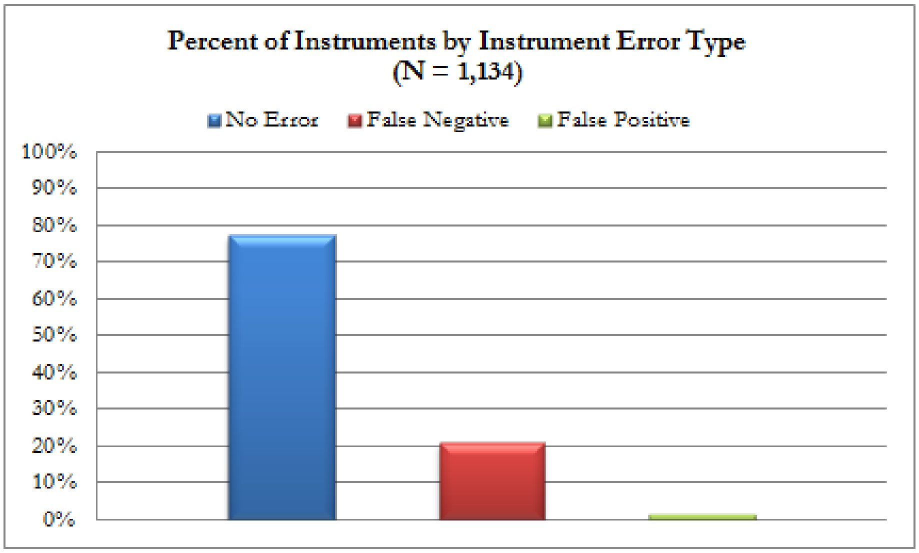

Figure 3.7: Error Rate in Total Instrument Sample by Error Type

3.3.3 Data Quality

3.3.3.1 Inconsistent Granularity

In addition to abstraction errors, there were a number of issues with the quality of data in this field. First, the field represents instruments at various levels of granularity. The types of entities captured in the TM field include instrument classes (i.e., unqualified instrument names); instruments; and instrument components, such as scales, subscales, indices and scores. Examples of entities of each type found in our dataset are listed in Table 3.1.

Table 3.1: Multi-Granularity Examples

PsycINFO® TM Entry	Entity Type	Instrument
Barratt Impulsiveness Scale	Instrument Class	BIS-10[1], BIS-11A[2], BIS-11[3]
Symptom Checklist 90 Revised	Instrument	SCL-90-R[4]
Symptom Checklist-90 Revised-Interpersonal Sensitivity Subscale	Scale	SCL-90-R[4]
Isolation Intolerance Subscale	Subscale	ZKPQ[5] Sociability Scale
Global Severity Index	Index	SCL-90-R[6]
EuroQol-5D Utility Score	Score	EuroQol-5D

[1]Barratt Impulsiveness Scale Version 10; [2]Barratt Impulsiveness Scale Version 11A; [3]Barratt Impulsiveness Scale Version 11; [4]Symptom Checklist 90 Revised; [5]Zuckerman Kuhlman Personality Questionnaire

In our sample of 203 publications and 1,134 associated TM entries, 41 (4%) TM entries represented instrument components. An informal analysis in which we queried the larger database for all entries in the TM field containing the string "symptom checklist" and three known indices of two versions of the instrument ("global severity index", "general severity index", and "positive symptom distress index") returned a list of 242 TM entries. Of these entries, 63% were instrument entities, 20% were index entities, and 17% were scale entities. Variation in granularity is not problematic if explicitly communicated, so users are able to effectively query the field. One aspect of variability that *is* problematic relates to variability in the way in which instrument components are represented in the field. In some cases, the component name is captured and qualified by the parent instrument (e.g., “Symptom Checklist 90 Revised Global Severity Index”). In other cases, the component is not qualified (e.g., “Global Severity Index”). This becomes problematic when the unqualified component captured shares a name with components of other instruments (e.g., “borderline scale” or “anxiety scale”), and therefore cannot be disambiguated in the search. In our sample, 21 (51%) of the TM entries representing instrument components were unqualified.

3.3.3.2 Lack of Standardization

A second concern we had with data in the TM field is related to lack of standardization. There is no controlled vocabulary for the field, and abstractors do not have access to a formal “authority file”[62] (i.e., a controlled vocabulary, or master list of standardized instrument names from which to select values). Consequently, names of entities captured in the TM field are idiosyncratic and highly variable. Spelling errors and idiosyncratic use of punctuation are pervasive. In some cases, the entity captured is the verbatim name provided by the author in the publication. In other cases, there appears to be an effort to represent the instrument using some sort of standard naming convention (as evidenced by use of a more standardized term than that used by the author in the publication).

Table 3.2: Examples of Non Standardization

Standardization Issue	PsycINFO® TM Entry
Use of non-standard term (i.e., acronym)	BIS
Use of incorrect term (i.e., “impulsivity” versus “impulsiveness”; “Barrett” versus “Barratt”; “II” versus “11”)	Barratt Impulsivity Scale Barratt Impulsivity Scale-11th Version Barratt Impulsivity Scale-II Barrett Impulsiveness Scale Barratt Impulsiveness Scale-Version II
Superfluous use of punctuation	Barratt Impulsiveness Scale. Barratt Impulsiveness Scale-11. Barratt Impulsiveness Scale, Version 11.
Inconsistent use of qualifier delimiter	Barratt Impulsiveness Scale-version 11 Barratt Impulsiveness Scale--version 11 Barratt Impulsiveness Scale, Version 11 Barratt Impulsiveness Scale 11
Inconsistent approach to qualifying instrument name	Barratt Impulsivity Scale-11th Version Barratt Impulsiveness Scale, Version 11 Barratt Impulsiveness Scale 11

3.4 Discussion

Results of our analyses reveal that publications in which standard psychological assessment instruments are used make up a substantial proportion of all peer-reviewed publications in the behavioral sciences and mental health domain. Psychometric instruments are, by definition, the foundation upon which virtually all empirical knowledge in this domain is based. Consequently, the ability to access publications in which specific psychometric instruments are used is essential to making a number of clinical and research decisions related to selection, use and interpretation of instrument results.

Currently, PsycINFO® is the only product on the market designed to provide comprehensive and near real-time access to the information required to make these decisions. However, our results indicate that this field accurately captures fewer than 80% of instruments used in in the published studies. For a user searching for a small set of studies in which a particular instrument was used, the consequences are minimal. However, for a user in search of comprehensive information about one or more specific instruments that can be used to make clinical or research protocol decisions, or for the reviewer attempting to perform a systematic review of the literature related to a particular instrument, the consequences are far more significant. In fact, the high error rate found in this study suggests that the poor quality of data in the TM field may undermine the immense and unique potential value of this field.

3.4.1 Recommendations

3.4.1.1 An Explicit Abstraction Algorithm and Automated Abstraction Process

The number of missed false negative errors by our own raters during manual abstraction suggests that a manual, human abstraction process is inherently biased towards in a high number of false negative errors. Our review of these errors suggests that when instruments are described by authors in multiple sections of a publication, or not described in detail, human abstractors will frequently miss them. Consequently, this process would likely benefit from development of a partially automated process.

Whether a manual or automated abstraction process is used, an explicit, formal algorithm for abstracting instruments should be defined and implemented. The algorithm must include explicit rules about the level of granularity at which TM entities are to be abstracted. For example, the decision must be made regarding whether the field captures instrument class, instrument, or instrument component data. If the decision is made to allow entities at all levels of granularity, the algorithm must specify the decision rules for level of granularity. For example, will the preferred level of granularity be the instrument? Will the instrument class only be allowed when the publication does not provide sufficient information to identify the specific instrument? When the publication describes the use of specific scales, subscales, indices, or scores, will these be abstracted? What will determine when instrument components will be abstracted? Will components be abstracted only when the author specifically identifies the component in the methods section, or reports metrics at the component level?

Finally, if the TM field allows instrument components, how will the data be represented? Will an entry be made for both the parent instrument and the component, or will the component be represented by qualifying the instrument name with the component name? For example, in a study that uses only the "Borderline Features Scale" of the "Personality Assessment Inventory" (PAI), the decision will need to be made whether to create a single TM field entry (e.g., "Personality Assessment Inventory, Borderline Features Scale"), or two entries (e.g., "Personality Assessment Inventory" and either "Personality Assessment Inventory, Borderline Features Scale" or "Borderline Features Scale"). The initial option seems to make the most sense from a usability perspective. However, all three approaches are evident in the TM data set analyzed in the current study.

3.4.1.2 A Controlled Vocabulary of Standard, Normalized Instrument Names

Creation and use of an "authority file" or controlled vocabulary of standard, normalized instrument names can be used to standardize entities captured in the TM field. This will resolve a number of issues related to the extensive variability in values used in the field. Table 3.3 presents a list of standard, normalized instrument names (i.e., fully specified terms) from the instrument database created for this study. We chose to use the base instrument name along with a small number of keywords and keyword qualifiers to construct instrument version names. We used the terms 'version', 'edition', 'translated', and 'translation' as keywords. We used the terms 'original', 'modified', and 'reconstructed' as qualifiers for keyword *version*; and a standard language list as qualifiers for keyword *translation*. We qualified *edition* with an integer value representing the number of items in the instrument followed by the term 'item'; and qualified *translated* with the year in which the translation was published. Finally, we further refined *version* qualifiers by appending the publication year. For modified versions, we decided to append the names of the first two authors immediately following publication year.

Table 3.3: Standard, Normalized Naming Convention

Fully Specified Name
Hamilton Rating Scale for Depression
Hamilton Rating Scale for Depression Version Original 1960 Edition 17 Item
Hamilton Rating Scale for Depression Version Original 1960 Edition 21 Item
Hamilton Rating Scale for Depression Version Original 1960 Edition 21 Item Translated 1996 Translation German
Hamilton Rating Scale for Depression Version Modified 1976 Guy Edition 24 Item
Hamilton Rating Scale for Depression Version Modified 1979 Hedlund Vieweg Edition 21 Item
Hamilton Rating Scale for Depression Version Modified 1985 Miller Bishop Edition 21 Item
Hamilton Rating Scale for Depression Version Modified 1987 Rabkin Klein Edition 21 Item
Hamilton Rating Scale for Depression Version Reconstructed 1987 Riskind Beck Edition 21 Item

Some mechanism for specifying instrument version attributes should be developed, as many instruments have more than one available version. An analysis of the data in our instrument database, for example, revealed that more than one-third of all instruments had more than one version, and 6% had more than five (Figure 3.8).

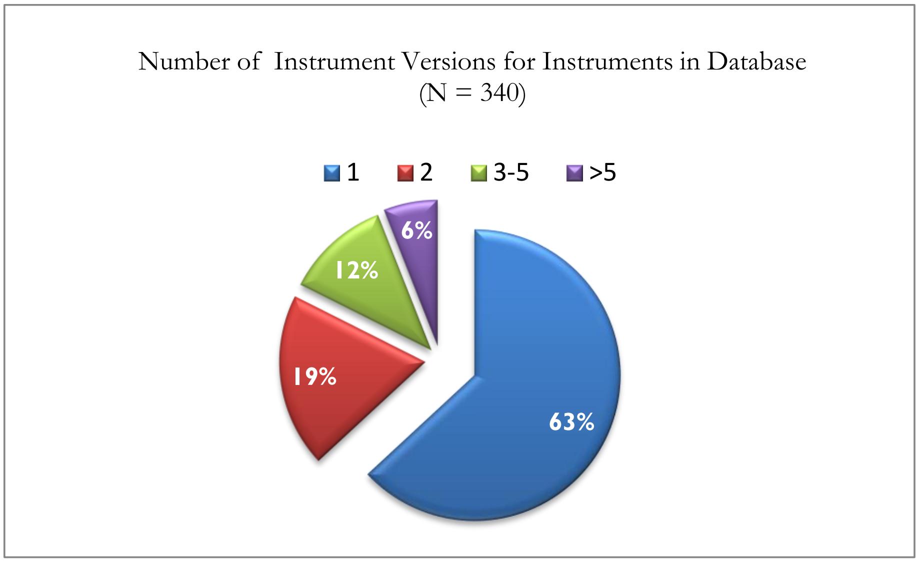

Figure 3.8: Proportion of Instruments by Number of Available Versions

Using a fully specified name (FSN) structured in this manner was useful for mapping instruments in our database, because it allowed us to see (in a single line item) the information elements commonly used in publications to describe the specific instrument used. This improved the accuracy of our mapping process compared to our original method (which was to use a commonly used term, and search our database for instrument details when we were unsure of the version). The FSN, however, is not a user-friendly expression, and therefore not ideally suited to the TM field. A reasonable solution, and one used ubiquitously in the terminology development field, is to select a preferred term, and map this term to the FSN. This approach is depicted in table 4 below. This type of approach could provide abstractors with the information necessary to accurately abstract instruments, while simultaneously providing a standardized, user-friendly term for the TM field.

Table 3.4: Standard Normalized Naming Convention with Preferred Terms

Fully Specified Name	Preferred Term
Beck Depression Inventory	Beck Depression Inventory
Beck Depression Inventory Version Original 1961 Edition 21 Item	Beck Depression Inventory Version I
Beck Depression Inventory Version Revised 1978 Edition 21 Item	Beck Depression Inventory Version IA
Beck Depression Inventory Version Revised 1996 Edition 21 Item	Beck Depression Inventory Version II
Beck Depression Inventory Version Revised 1997 Edition 7 Item	Beck Depression Inventory Primary Care

3.4.1.3 Mechanism for Error Detection during Abstraction

Abstraction errors can be addressed by providing abstractors with a tool for verifying that author-supplied instrument information is consistent. For example, the abstractor can enter the instruments details provided by the author in the methods section of the paper (e.g., instrument name, author, item count, version number, etc.). The tool can provide the standardized instrument name, if a complete match is found.

Alternately, it could provide a list of candidate instruments, along with information about inconsistencies between data reported by the author.

3.5 Conclusions

Publications indexed by PsycINFO® are a rich source of information about available psychological assessments and psychometric instruments. However, the ability to access this data is contingent upon the quality and completeness of data in the TM field of this database. We empirically evaluated data in this field, and found that 57% of publications in our sample contained at least one abstraction error. Most of these errors were false negative errors. These findings make it difficult to use the TM field of PsycINFO® as a reliable source of comprehensive information about instruments.

The current study has several limitations. First, all data upon which the current analysis was based were related to a single mental health condition (BPD). While the publications used in this analysis spanned a broad range of journals, topics, and time periods, it is possible that the publications used in our analysis are not representative of the entire set of publications indexed by PsycINFO®. Caution should be taken in generalizing the results of this study to the entire set of publications or TM entries in PsycINFO®.

Nonetheless, we recommend that a systematic effort be made to correct existing errors, and standardize entries in this field. We recommend that a systematic and comprehensive analysis of the quality of the data in this field be undertaken by the vendor because of the immense value of this data. It is our hope that the results of this analysis can be used to guide the development of a systematic process for abstracting and coding data for this field. We also recommend that a standard naming convention and algorithm for normalizing instrument names be developed. This naming convention will need to address the creation of both fully specified names and preferred terms. Ideally, this work will take place in the context of related work, and in collaboration with terminologists working with SNOMED CT[24], LOINC[63], MeSH, and the Thesaurus of Psychological Index Terms[34].

Finally, the use of natural language processing (NLP) techniques should also be explored. NLP has proven to be a highly effective technique for reliably extracting data from a variety of data sources in the healthcare enterprise[64,65-67]. There is no reason to believe that identification and extraction of instrument data in peer reviewed publications present a unique challenge that cannot benefit from the application of these techniques.

CHAPTER 4

PSYCHOLOGICAL ASSESSMENT INSTRUMENTS: A COVERAGE ANALYSIS USING SNOMED CT, LOINC AND CDISC'S QS TERMINOLOGY

Psychometric instruments, inventories, surveys, and questionnaires are widely accepted tools in the field of behavioral health. They are used extensively in primary and clinical research, patient care, quality measurement, and payor oversight. To accurately capture and communicate instrument-related activities and results in electronic systems, existing healthcare standards must be capable of representing the full range of psychometric instruments used in research and clinical care. Several terminologies and controlled vocabularies contain representations of psychological instruments. While a handful of studies have assessed the representational adequacy of terminologies in this domain, no study to date has assessed content coverage. The current study was designed to fill this gap. Using a sample of 63 commonly used instruments, we found no concept in any of the three terminologies evaluated for more than half of all instruments. Of the three terminologies studied, SNOMED CT (Standard Nomenclature of Medicine – Clinical Terms) had the greatest breadth, but least granular coverage of all systems. While SNOMED CT contained concepts for over one third (36%) of the instrument classes in this sample, only 11% of the actual instruments were represented in SNOMED CT. LOINC (Logical Observation Identifiers, Names, and Codes), on the other hand, was able to represent instruments with the greatest level of granularity of the three terminologies. However, LOINC had the poorest coverage, covering fewer than 8% of the instruments in our sample. Given that instruments selected for this study were selected on the basis of their status as gold standard measures for conditions most likely to present in clinical settings, we believe these results overestimate the actual coverage provided by these terminologies. The results of this study demonstrate significant gaps in existing healthcare terminologies vis-à-vis psychological instruments and instrument-related procedures. Based on these findings, we recommend that systematic efforts be made to enhance standard healthcare terminologies to provide better coverage of this domain.

4.1 Introduction

Psychometric instruments, inventories, surveys, and questionnaires are widely accepted tools in the field of behavioral health[28]. In patient care, they are used to screen for and diagnose mental health conditions. They are used to obtain information about the nature and severity of symptoms, and to select appropriate treatments. They are also used to monitor patients' response to treatment, and to detect potential adverse reactions[10]. In primary research, instruments are used to screen participants, and to operationalize study variables. In clinical research, instruments are used to assess the efficacy and effectiveness of specific therapeutic programs and to compare the relative effectiveness of two or more interventions. Finally, in quality improvement and general oversight, instruments are used to identify gaps in quality and guide performance improvement programs.

As a central source of objective data upon which all evidence-based practice in this domain is founded, the ability to capture, aggregate, and share instrument-related information is a high priority for both researchers and clinicians[45,42,68]. While a number of studies exist describing the adequacy of representations in this domain, no study to date has evaluated the coverage of instruments in any healthcare terminology. The current study was undertaken to fill this gap. Specifically, this study evaluates three clinical terminologies available to users of electronic information systems for capturing structured information related psychological assessments. The first two terminologies, SNOMED CT and LOINC, are designed for use in the clinical domain. The third, QS Terminology, is designed specifically for use in medical and behavioral research. Each of these three terminologies allows for the structuring of data related to a slightly different set of assessment-related elements and procedures. All three, however, contain representations corresponding to the instrument-level entity.

The primary aim of the current study is to evaluate the coverage of these three terminologies for psychological assessment instruments currently used in research and clinical practice. A secondary aim of this study is to evaluate the consistency with which these terminologies currently represent psychological instruments.

4.2 Background

In 2006, the Institute of Medicine (IOM) published an extensive report arguing that current efforts to improve the quality of healthcare must be extended to behavioral health[3]. A key recommendation of this report is that the health information infrastructure be enhanced to accommodate the needs of behavioral healthcare[3]. In this report, the IOM specifically identifies the creation of standard terminologies for all commonly used interventions and assessment methods as a key element in building this infrastructure. Moreover, in 2009, the United States Government passed legislation requiring healthcare providers seeking reimbursement for federally funded services to use electronic health records that comply with minimum standards. Among these standards are the requirement that all assessment data submitted to key external organizations be coded using a pre-specified set of standard identifiers.

Currently, there are three standard terminologies available for capturing, aggregating and communicating psychological assessment data obtained in clinical and research processes. These terminologies are SNOMED CT, LOINC, and CDISC's Questionnaire (QS) Terminology. SNOMED CT, the Standard Terminology for Medicine – Clinical Terms, is a clinical terminology used for capturing, aggregating, analyzing and sharing granular clinical data that span virtually the entire range of healthcare. SNOMED CT contains more than 300,000[69] concepts across a wide range of domains including diseases, anatomical structures, clinical findings, procedures, social contexts, devices and biological substances[70]. LOINC® (Logical Observations, Identifiers, Names and Codes) is a terminology consisting of a set of universal codes for laboratory tests and other types of clinical observations and measures[71]. In 2000, LOINC introduced the SURVEY class to accommodate clinical assessments[45], which includes questionnaires, surveys, and other form-based assessment instruments. Currently, LOINC includes entries for thirteen psychological assessment instruments, six of which are modeled as components of the mental health survey class (SURVEY.MLTHLTH).

While SNOMED and LOINC are designed specifically for use in clinical information systems used at the point of care, the third terminology, Questionnaire (QS) Terminology, is designed specifically for use in clinical research. QS Terminology was developed by the Clinical Data Interchange Standards Consortium (CDISC), a non-profit Standards Development Organization (SDO) whose primary objective is to support the capture and exchange of data in "all types of medical research"[72]. QS Terminology is one standard in a suite of several that are designed to support structured data collection at all stages of the research process, from protocol development to analysis and reporting of results. These standards are specifically designed to support the acquisition, exchange, and submission of clinical research data[72].

Both SNOMED CT and LOINC are endorsed by the United States government for use in electronic health records (EHRs). As part of the HITECH (Health Information Technology for Economic and Clinical Health) Act of 2009[6], the government mandated that electronic health record systems used by healthcare delivery organizations meet requirements for the 'meaningful use' of health information. To meet this requirement, EHRs must code relevant clinical data using approved terminologies. SNOMED CT and LOINC are two terminologies specifically designated by the US government as acceptable for coding granular clinical data (i.e., data that cannot be represented using standard diagnostic and procedural coding systems). While both terminologies contain concepts for coding psychological assessments, the United States' national strategy is to represent such instruments, along with all other types of measures, in LOINC[48].

Very little research, to date, has focused on the role of standard terminologies for representing psychological assessments. What has been published in this domain has focused primarily on the adequacy of representations in existing terminologies. This work has focused largely on LOINC. Four key studies have been performed addressing standardized assessments in LOINC. Two of these studies address modifications to LOINC necessary to accommodate assessment instruments. The first of these studies was performed by Bakken et al[45] in 2000. This paper introduced a new LOINC class to accommodate psychological assessments, and proposed extensions to the six core axes used to represent entities in LOINC[45]. The second study, performed by White et al[42], addressed the important issue of instrument versioning as it relates to psychometric properties of instruments and its impact on representation of assessments in LOINC. These authors proposed four modifications to LOINC they believed necessary to support the accurate identification and disambiguation of assessment instruments. In addition, two of the four publications addressed LOINC content specifically as it relates to standardized assessments[42,45]. Neither study focused specifically on psychological assessment instruments, nor did either study address coverage. To date, no study has systematically evaluated the coverage of instruments in currently available terminology products. Furthermore, our literature search did not return a single article addressing either the representation, or coverage, of psychological assessments in SNOMED CT.

4.3 Methods

4.3.1 Datasources

Data for the current study came from healthcare web sites and three standards development organizations (SDOs). The web sites used for the current study were the National Institute of Mental Health[25] (NIMH), and those of several organizations dedicated to the research and treatment of mental health conditions. Organization web sites were selected as the primary data source for identifying both target mental health conditions and high quality, commonly used condition-specific instruments. We selected web sites as the primary data source because we believed them to the most valid and timely source for this information. The terminologies evaluated in this study are SNOMED CT version 2012.8.0270, LOINC version 2.40, and CDISC's QS Terminology (Figure 4.1).

Table 4.1: Terminology Datasources

Terminology	Organization	Target Information System	Mandate
LOINC	Regenstrief	Electronic Health Records	HITECH Meaningful Use
SNOMED CT	IHTSDO[1]	Electronic Health Records	HITECH Meaningful Use
QS Terminology	CDISC[2]	Clinical Research Information Systems	Industry Consensus Standard

[1]International Healthcare Terminology Standards Development Organization; [2]Clinical Data Interchange Standards Consortium

4.3.2 Procedure

Using the NIMH list of mental health conditions as a source, we identified sixteen disorders. To this list, we added four additional instrument categories we believed to be particularly relevant to mental health (Appendix C). For each condition or construct, we selected instruments based on recommendations provided by organizations dedicated to the study or treatment of the target condition. Our goal was to select a set of 3 to 4 instruments for each condition. This resulted in a master set of 63 instruments (Appendix D).

For the current study, we evaluate ten distinct information elements related to instruments. These elements, along with the algorithm used to code them, are depicted in Table 4.2. The primary unit of analysis for this study is the discrete assessment instrument. We define an assessment instrument as a named collection of observations and observation metrics. It is a concrete entity that is fully defined by its component elements (i.e., scales, subscales, items) and the methods used to generate an observation (i.e., item response scale and metric). Consequently, when changes are made to either the instrument components (e.g., observations are added, deleted, or substantially modified) or the method by which observations are evaluated and quantified (e.g., response scale changes), the collection is assumed to represent a new instrument.

We define an instrument class as the more general class of observations from which specific instruments (instrument versions) are derived. The instrument class can be thought of as the “base” instrument upon which one or more specific instrument versions are based. The ‘Beck Depression Inventory’ (BDI) is an example of an instrument class. While the base instrument is not simply the first instrument in a series of versions or revisions, many key instrument attributes (e.g., construct measured, target population, etc.) are defined by the first instrument. In this sense, the first instrument can be thought of as the class template from which future instances are derived.

Table 4.2: Instrument Information Elements and Coding Algorithms

Information Element	Coding Algorithm
Instrument Class	The terminology contains a concept for representing the instrument class.
Instrument Instance	The terminology contains a concept for representing the specific version of the instrument in the sample.
Instrument Scale	If items in the instrument are organized into named scales, the terminology contains a concept for representing at least one of the scales in the instrument.
Instrument Subscale	If items in the instrument are organized into scales, and one or more of the scales is organized into subscales, the terminology contains a concept for representing at least one of the subscales in the instrument.
Instrument Item	The terminology contains a concept for representing at least one item in the instrument.
Instrument Response Items	The terminology contains a concept for representing at least one item response in the instrument.
Instrument Response Values	The terminology contains a concept for representing at least one value for one item response in the instrument.
Instrument Score	The terminology contains a concept for representing at least one score generated by the instrument.
Instrument Administration	The terminology contains a concept for representing administration of the specific instrument.
Instrument-Related Findings	The terminology contains a concept for representing one or more findings specifically related to use of the instrument.

Using the example of the BDI, one class (i.e., BDI) and four instruments (i.e., BDI-I, BDI-IA, BDI-II, and BDI-PC) can be identified. The first three instruments represent a refinement (or adaptation) of the instrument over time. The fourth instrument, the BDI-PC, represents a modification that has been made for use in a new setting. The boundary between instrument classes can sometime become blurry, as when an instrument is adapted from a previous instance, and the new instance is sufficiently different from the original class that it becomes a new class. However, this case did not apply to any of the instruments in the current sample, and is therefore is not addressed further in the current analysis.

The next five information elements correspond to instrument components. These are instrument *scales*, *subscales*, *items*, *item responses*, and *item response values*. An instrument scale is defined as a named collection of observations within the instrument, and a subscale as a named collection of observations within a scale. Instrument items are defined as the discrete observations that are coded. Item responses are defined as the discrete set of responses that can be associated with the observation. Item response values are defined as the specific code (nominal or quantitative) that can be associated with the item response[31]. Instrument scores are defined as quantitative or qualitative scores derived from one or more scales, subscales or observations. For purposes of the current study, we evaluated only the score associated with the overall instrument (i.e., total or summary score).

In addition to concepts related to the physical instrument, we also assessed whether the terminology contained concepts related to instrument administration and instrument-specific findings. Instrument administration is defined as the activity during which the set of observations contained in the instrument are made. Instrument-specific findings are defined as clinical observations that are made on the basis of results obtained by administering the instrument (e.g., 'decrease in depression score', 'no change in suicidal ideation').

Using the ten information elements identified above, we coded each terminology according to whether or not it contained a corresponding concept for each instrument. This strictly binary method differs slightly from previous methodologies used in coverage studies. In a coverage analysis of clinical concepts found in medical records, for example, Chute and colleagues[73] used a dimensional model in which they rated each concept along a continuum representing closeness of match to the target concept. In the Chute study, each concept was scored on a scale of 0 (no match) to 2 (exact match). Wasserman and Wang[9] took a slightly different approach in their coverage analysis. Examining concepts required for implementing Computerized Provider Order Entry (CPOE), these authors used a 4-level coding scheme that essentially quantified the effort required to add the target concept to the terminology. In the current study, we chose to explicitly define the specific instrument-related information element we were targeting, and to code the corresponding concept in a strictly binary (i.e., present/not present) manner. We used this approach because we felt it provided an accurate assessment of the value of the terminology from the perspective of the end user. That is, an end user would either find a useable concept, or not.

Using this scheme, each of the three terminologies was searched for concepts corresponding to the ten target information elements for each of the 63 instruments in the sample. Interrater reliability was calculated based on a sample of 15 instruments (150 target concepts) coded by a second reviewer (TJA) and found to be adequate ($\kappa = 0.92$). A review of discrepancies between raters revealed that most occurred in distinguishing between *instrument instance* concepts and the *instrument class* concept in SNOMED CT. Interrater agreement for *instrument instance* and *instrument class* concepts was lower than overall agreement ($\kappa = 0.88$). Interrater agreement for SNOMED CT was also lower ($\kappa = 0.89$).

4.4 Results and Discussion

4.4.1 Terminology Coverage

Coverage of *instrument instance* concepts in all three terminologies was poor. These three terminologies covered an average of 5.6 (9.1%) *instrument instances* in our sample. Of the three terminologies evaluated, SNOMED CT had the best coverage (n = 7, 11.3%) and LOINC the poorest (n = 4, 6.5%).

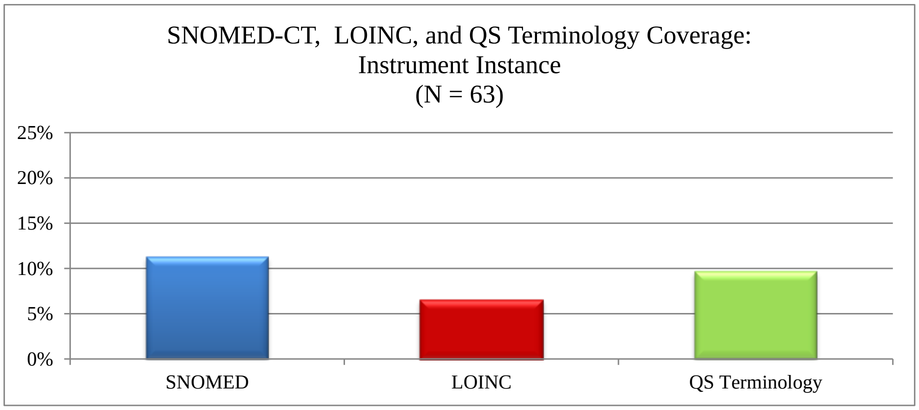

Figure 4.1: Instrument and Instrument Components Representation

A summary of concept coverage by terminology is presented in Table 4.3. SNOMED CT covered the largest number and widest range of overall instrument-related concepts, containing 48 concepts across five of the ten instrument concept categories. CDISC's QS Terminology contained the greatest number of *instrument item* concepts, but covered the smallest range of concepts. LOINC was the only terminology providing coverage of *instrument response* concepts. SNOMED CT was the only terminology providing coverage of *instrument administration* or *instrument findings* concepts (**Error! Reference source not found.**).

Table 4.3: Instrument Coverage by Terminology

Attribute	SNOMED CT	LOINC	QS Terminology
Instrument Class	22 (35%)	2 (3%)	-
Instrument Instance	7 (11%)	4 (6%)	6 (10%)
Instrument Scale	-	-	-
Instrument Subscale	-	-	-
Instrument Item	-	3 (5%)	6 (10%)
Instrument Response Items	-	3 (5%)	-
Instrument Response Values	-	-	-
Instrument Score	8 (13%)	1 (2%)	-
Instrument Administration	9 (15%)	-	-
Instrument Related Finding	1 (2%)	-	-

4.4.1.1 Instrument Class and Instance

While SNOMED CT contained the greatest number of instrument-related concepts in our sample, most of these concepts are not specific enough to be used to code or capture clinical data. Instead, these concepts are best used as organizing concepts under which missing child concepts can be modeled (Table 4.4). For example, 22 (46%) of the 48 concepts covered by SNOMED CT were *instrument class* concepts. Moreover, of the seven *instrument instance* concepts covered by SNOMED CT, only one was modeled with sufficient detail to ensure that the concept remains unambiguous with the introduction of future instrument versions. That is, the concept for only one instrument in the sample contained sufficient versioning information to use the instrument concept to unambiguously identify a particular instrument.

Table 4.4: SNOMED CT Coverage of Instrument Class and Not Instruments

Instrument Class	Missing SNOMED CT Instrument Concepts
beck depression inventory (assessment scale)	beck depression inventory version 1 (assessment scale)
	beck depression inventory version 1a (assessment scale)
	beck depression inventory version 2 (assessment scale)
	beck depression inventory primary care version (assessment scale)

4.4.1.2 Instrument Component

There was no coverage of *instrument scale* concepts in any of the three terminologies for any of the instruments in our sample. A detailed review of concepts in each terminology, however, confirmed that both LOINC and SNOMED are capable of representing this class of concepts. In fact, we found several instances of *instrument scale* concepts in SNOMED CT. Interestingly, CDISC's QS Terminology provided better coverage of instrument element concepts in our sample than did LOINC. LOINC, however, was the only terminology that provided any coverage of *item response* concepts.

4.4.1.3 Instrument Administration

Both LOINC and QS Terminology codes can be used to indicate that a particular instrument or instrument item was administered. However, neither terminology provides codes corresponding to the concept of having administered an instrument. SNOMED CT, on the other hand, contains discrete codes for instrument administration. These codes are modeled in SNOMED CT as child concepts of concepts in the 'procedure by method' and 'provider-specific procedure' hierarchies. Both of these hierarchies are in turn children of the top level 'procedure' hierarchy, which represents activities performed during the routine delivery of healthcare[24]. Ten (77%) of the 13 instruments with corresponding administration procedure concepts in SNOMED CT were modeled as child concepts of 'assessment using assessment scale', and three (23%) were modeled as child concepts of 'psychologic test' (Table 4.5).

Table 4.5: SNOMED CT Representation of Instrument Administration Procedure

SNOMED CT Hierarchy	Sample Concepts
Provider-specific procedure	Minnesota Multiphasic Personality Inventory (MMPI-2)
Psychologic evaluation or test procedure	Rorschach
Psychologic test	Thematic Apperception Test (TAT)
Procedure by method	Beck Depression Inventory (BDI)
Evaluation procedure	Eating Disorder Examination (EDE)
Assessment using assessment scale	Eating Disorders Inventory (EDI)
	Edinburgh Postnatal Depression Scale (EPDS)
	Mini-Mental State Examination (MMSE)
	Penn State Worry Questionnaire (PSWQ)
	Psychotic Symptom Rating Scale (PSYRATS)

Interestingly, while SNOMED CT included concepts representing the administration of 17 distinct instruments under the concept ‘psychologic test’, it contained a concept representing the corresponding instrument for only 8 (47%) of the instrument administration procedures modeled. Finally, SNOMED CT was unique among the three terminologies in our study in that it was the only terminology that included any concept related to interpretation of instrument results. Admittedly, we found only two qualifying concepts, and both referred to the same instrument (i.e., the ‘Edinburgh postnatal depression scale (assessment scale)’. These concepts were ‘decline in Edinburgh postnatal depression scale score (finding)’, and ‘decline in Edinburgh postnatal depression scale score at 8 months (finding)’.

4.5 Conclusion

The results of this study demonstrate significant gaps in existing healthcare terminologies for psychological instruments and instrument-related activities. In this study, we evaluated three standard terminologies and found no concept in any of the three terminologies for more than half of all instruments evaluated. On average, these three terminologies covered fewer than 10% of the instrument instances in our sample, with a range of 7% (LOINC) to 11% (SNOMED CT).

The primary limitation of this study is that many of the instruments included in our sample are copyrighted and/or commercial instruments. Consequently, they may not exist in standard terminologies due to an unwillingness of the copyright holder to release the copyright. Similarly, the results of this study cannot be generalized to all standardized assessment instruments. Assessment instruments cover a broad range of healthcare domains, beyond just mental health. The instruments selected for the current study were all psychometric instruments. It is quite possible that coverage of standardized assessments for another domain (e.g., oncology or neuropsychology) would yield a different result.

However, given that instruments selected for this study were selected on the basis of their status as gold standard measures for conditions most likely to present in clinical settings, we believe these results may underestimate the actual coverage of psychometric instruments provided by current healthcare terminologies. Based on these findings, we recommend that systematic efforts be made to enhance standard terminologies to provide better coverage of this domain. Specifically, instruments currently used in research and routine clinical care should be identified, and inclusion of the identified instruments in at least one standard terminology made a priority.

.

CHAPTER 5
CONCLUSION

Developing high quality information systems capable of supporting research and clinical care in behavioral health requires the existence of robust clinical terminologies. These terminologies must be capable of representing the same breadth and depth of concepts in the behavioral domain that we demand of terminologies in the medical domain. The results of this study demonstrate that significant gaps exist in terminologies relative to behavioral health.

Focusing specifically on psychological assessment instruments and their role in healthcare, we assessed the extent to which existing healthcare terminologies can be used to capture, code, aggregate, and retrieve information in this domain. In the first study, we reviewed and harmonize representations of psychological instruments found in a current healthcare terminologies, ontologies, and information products. We identified a comprehensive set of 86 instrument-related information elements, and organized these elements in eight classes of information. We then empirically assessed the form (i.e., explicit, implicit) and consistency with which these representations are implemented.

In this analysis, we identified several gaps in terminologies. Most significantly, we found that none of the terminologies we examined is capable of fully representing instrument identifiers. Consequently, there is currently no standard way to unambiguously identify specific psychological assessment instruments in the healthcare enterprise. Moreover, no terminology examined is capable of representing information related to the administration or scientific properties of these instruments.

In the second study, we assessed the accuracy and completeness of data contained in PsycINFO®. PsycINFO® is the most comprehensive database currently available for accessing information about in-vivo instrument performance. Using a sample of 203 publications, we empirically assessed the accuracy and quality of data in the *Tests and Measures* (TM) field of this database. We found that than 57% of the publications in our sample contained at least one abstraction error. Based on these findings, we concluded that PsycINFO®, and the TM field in particular, cannot be used as reliably comprehensive source of information about psychological instruments.

In the third and final study, we evaluated the extent to which existing terminologies cover a set of instruments commonly used in behavioral health today. We evaluated three terminologies most relevant to representing psychometric instruments and other psychological assessments: CDISC's QS Terminology[21,22], LOINC[23], and SNOMED CT[24]. Using a list of mental health conditions and topics listed on the National Institute for Mental Health (NIMH) web site[25], we assessed how well each terminology covered ten discrete concepts related to each instrument. We found no concept in any of the three terminologies for more than half of all instruments evaluated. On average, these three terminologies covered fewer than 10% of the instruments in our sample, with a range of 7% (LOINC) to 11% (SNOMED CT).

Taken together, the results of these three studies demonstrate significant gaps in healthcare terminologies relative to psychological assessment instruments. The ability to represent the essential components of, and activities related to, psychometric instruments is essential to the provision of high quality care in behavioral health. Psychometric instruments are, by definition, the foundation upon which all empirical knowledge in this domain is based. These instruments play a central role in shaping current understanding of both clinically relevant phenomena and specific mental health conditions. Moreover, results obtained using instruments are the grounds upon which clinical practice in this domain is designated as "evidence based practice".

These gaps in terminologies have serious implications for the quality of behavioral health care. The inability to represent psychological assessment instruments using structured, unambiguous codes hampers the application of HIT in the behavioral health domain. Moreover, it undermines efforts to improve quality, as those features shown to improve quality (e.g., computerized provider order entry, clinical decision support) all depend on the ability of EHRs to code clinical data in a structured and standardized format[7,8,9,10].

It is clear from this study, that it is time to embrace the recommendations of the 2006 IOM Report. It is time to aggressively work to extend the findings of earlier IOM reports to the domain of behavioral health. Only by focusing first on standard terminologies – the foundation upon which all HIT is built –can we continue to expand the NHII to include behavioral health.

BIBLIOGRAPHY

1. Institute of Medicine. To err is human : building a safer health system. Washington, D.C.: National Academy Press; 2000.

2. Institute of Medicine. Crossing the quality chasm : a new health system for the 21st century. Washington, D.C.: National Academy Press; 2001.

3. Institute of Medicine. Improving the quality of health care for mental and substance-use conditions. Washington, DC: National Academies Press; 2006.

4. US Congress. House. Health Information Technology. American Recovery and Reinvestment Act of 2009, HR 1, 111th Congress 2009;Section A, Title XIII.

5. US Congress. House. American Recovery and Reinvestment Act of 2009. HR 1, 111th Congress 2009.

6. Meaningful Use. U.S. Department of Health and Human Services, 2012. (Accessed November 18, 2012, at http://www.healthit.gov/policy-researchers-implementers/meaningful-use.)

7. Bales ME, Lussier YA, Johnson SB. Topological analysis of large-scale biomedical terminology structures. J Am Med Inform Assoc 2007;14:788-97.

8. Elkin PL, Brown SH, Chute CG. Guideline for health informatics: controlled health vocabularies--vocabulary structure and high-level indicators. Stud Health Technol Inform 2001;84:191-5.

9. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annu Symp Proc 2003:699-703.

10. Lambert MJ. Prevention of treatment failure : the use of measuring, monitoring, and feedback in clinical practice. 1st ed. Washington, DC: American Psychological Association; 2010.

11. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (4th ed., text rev.). Washington, DC: Author; 2000.

12. DSM-5 Overview: The Future Manual. 2012. (Accessed November 24, 2012, at http://www.dsm5.org/about/Pages/DSMVOverview.aspx.)

13. National Center for Health Statistics. The International Classification of Diseases, Tenth revision, Clinical Modification (ICD-10-CM). New York, NY: Author; 2013.

14. National Center for Health Statistics. The International Classification of Diseases, Tenth revision, Procedural Classification (ICD-10-PC). New York, NY: Author; 2013.

15. American Medical Association. Current Procedural Terminology (CPT): American Medical Association Press; 2012.

16. World Health Organization. International Classification of Functioning, Disability and Health (ICF): Author; 2001.

17. The National Institute of Mental Health Strategic Plan. National Institute of Health, 2008. (Accessed November 23, 2012, at http://www.nimh.nih.gov/about/strategic-planning-reports/index.shtml.)

18. Research Domain Criteria (RDoC). 2012. (Accessed November 23, 2012, at http://www.nimh.nih.gov/research-funding/rdoc/index.shtml.)

19. Regier DA, Narrow WE, Kuhl EA, Kupfer DJ, eds. The conceptual evolution of DSM-5. Washington DC: American Psychiatric Publishing, Inc.; 2011.

20. Millon T, Krueger RF, Simonsen E. Contemporary directions in psychopathology: Scientific foundations of the DSM-V and ICD-11. In: Contemporary directions in psychopathology: Scientific foundations of the DSM-V and ICD-11. New York, NY: Guilford Press; US; 2010:xvii, 622.

21. CDISC. 2012. (Accessed October 4, 2012, 2012, at http://www.cdisc.org/.)

22. Terminology Resources: NCI Enterprise Vocabulary Services (EVS), Dictionaries, FedMed, FDA, CDISC, and NCPDP Terminology. 2012. (Accessed October 4, 2012, at http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/cdisc.)

23. Regenstrief Institute. Logical Observation Identifiers Names and Codes (LOINC®). In. Indiana IN: Regenstrief Institute, Inc.; 2012.

24. IHTSDO. Standard Nomenclature of Medicine-Clinical Terms. International Release January 2011.

25. Mental Health Topics. 2012. (Accessed August 12, 2012, at http://www.nimh.nih.gov/index.shtml.)

26. Lee YT. Information Modeling: From Design to Implementation. In: Saadat SNaM, editor. Proceedings of the Second World Manufacturing Congress; 1999; Canada/Switzerland: International Computer Science Conventions; 1999. p. 315-21.

27. W S. Frequently Performed Psychological Tests. In: Walker HK HW, Hurst JW, ed. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd ed. Boston: Butterworths; 1990.

28. White GW, Jellinek MS, Murphy JM. The use of rating scales to measure outcomes in child psychiatry and mental health. In: Baer L, Blais MA, eds. Handbook of clinical rating scales and assessment in psychiatry and mental health. New York, NY: Humana Press; 2010:175-94.

29. National Library of Medicine. In: PubMed Database. Washington DC: National Library of Medicine; 2012.

30. American Psychological Association. PsycINFO®.

31. Poldrack R. Cogntive Atlas: A Collaboratively Developed Cognitive Science Ontology Contributor's Guide 2009. In: University of Texas at Austin; 2009.

32. Russell A. Poldrack AK, Donald Kalar, Eric Miller, Christian Seppa, Yolanda Gil, D. Stott Parker, Fred W. Sabb, Robert M. Bilder. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. Frontiers in neuroinformatics 2011;5:11.

33. Bénédicte BATRANCOURT MD, Bernard GIBAUD, Gilles KASSEL. A core ontology of instruments used for neurological, behavioral and cognitive assessments. In: Conference on Formal Ontology in Information Systems; 2010; The Netherlands: IOS Press; 2010.

34. Tuleya LG. Thesaurus of Psychological Index Terms. 11th ed. Washington, DC: American Psychological Association; 2007.

35. Educational Testing Service. TestLink Database. In: Educational Testing Service; 2012.

36. Behavioral Measurement Database Services. Health and Psychosocial Instruments. In; 2012.

37. Buros Institute of Mental Measurements. Mental Measures Yearbook with Tests in Print. Available from http://buros.unl.edu/buros/.

38. American Psychological Association. PsycTESTS®. In. Washington DC: American Psychological Association; 2012.

39. Finding Psychological and Educational Tests. Regents of the University of Minnesota, 2010. (Accessed 7/12/2012, at http://www.lib.umn.edu/libdata/page.phtml?page_id=2415)

40. The National Center for Biomedical Ontology. BioPortal. http://bioportal.bioontology.org/.

41. National Library of Medicine. UMLS. http://www.nlm.nih.gov/research/umls/.

42. White TM, Hauan MJ. Extending the LOINC conceptual schema to support standardized assessment instruments. J Am Med Inform Assoc 2002;9:586-99.

43. Meyer GJ, Kurtz JE. Advancing personality assessment terminology: time to retire "objective" and "projective" as personality test descriptors. J Pers Assess 2006;87:223-5.

44. Bornstein RF. Toward a process-based framework for classifying personality tests: comment on Meyer and Kurtz (2006). J Pers Assess 2007;89:202-7.

45. 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. Journal of the American Medical Informatics Association : JAMIA 2000;7:529-38.

46. Vacha-Haase T, Kogan LR, Thompson B. Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. Educational and Psychological Measurement 2000;60:509-22.

47. Crocker L, Algina J. Introduction to Classical and Modern Test Theory. Mason, Ohio: Cengage Learning; 2008.

48. Humphreys B. Personal Communication.

49. Vacha-Haase T, Ness C, Nilsson J, Reetz D. Practices regarding reporting of reliability coefficients: A review of three journals. Journal of Experimental Education 1999;67:335-41.

50. Educational Testing Service. Tests in Print Database.

51. Goldman BA, Mitchell DF, eds. Directory of Unpublished Experimental Measures. Dubuque, IA: American Psychological Association; 2007.

52. K. T. Chun SC, & J. R. P. French, Jr. , ed. Measures for Psychological Assessment: A Guide to 3,000 Original Sources and Their Applications. Ann Arbor, MI: Institute for Social Research; 1975.

53. McDowell I. Measuring health: a guide to rating scales and questionnaires. New York, New York: Oxford Univserity Press; 2006.

54. Baer L, Blais MA, eds. Handbook of clinical rating scales and assessment in psychiatry and mental health. New York NY: Humana Press; 2010.

55. American Psychological Association. PsycINFO® Database.

56. American Psychological Association. APA Databases.

57. APA Databases: PsycINFO® Homepage. American Psychological Association,, 2012. (Accessed November 9, 2012, at http://www.apa.org/pubs/databases/psycinfo/index.aspx.)

58. American Psychological Association. June 2004 reload of the PsycINFO® database. PsycINFO® News 2004;24:1-3.

59. Paris J. Borderline personality disorder. In: Oxford textbook of psychopathology. 2nd ed. New York, NY: Oxford University Press; US; 2009:723-37.

60. Weinberg A, Klonsky ED, Hajcak G. Autonomic impairment in borderline personality disorder: a laboratory investigation. Brain Cogn 2009;71:279-86.

61. Roepke S, Schroder-Abe M, Schutz A, et al. Dialectic behavioural therapy has an impact on self-concept clarity and facets of self-esteem in women with borderline personality disorder. Clin Psychol Psychother 2011;18:148-58.

62. Hillson S. Personal Communication.

63. 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;5:276-92.

64. Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. J Am Med Inform Assoc 2012.

65. Thessen AE, Cui H, Mozzherin D. Applications of natural language processing in biodiversity science. Adv Bioinformatics 2012;2012:391574.

66. Danforth KN, Early MI, Ngan S, Kosco AE, Zheng C, Gould MK. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. J Thorac Oncol 2012;7:1257-62.

67. Workman TE, Stoddart JM. Rethinking information delivery: using a natural language processing application for point-of-care data discovery. J Med Libr Assoc 2012;100:113-20.

68. Vreeman DJ, McDonald CJ, Huff SM. Representing Patient Assessments in LOINC(R). AMIA Annu Symp Proc 2010;2010:832-6.

69. SNOMED CT Components. 2012. (Accessed November 18, at http://www.ihtsdo.org/SNOMED CT/SNOMED CT0/SNOMED CT-components/.)

70. International Healthcare Terminology Standards Development Organization. SNOMED CT User Guide. 2011.

71. Vreeman DJ, McDonald CJ, Huff SM. LOINC(R) - A Universal Catalog of Individual Clinical Observations and Uniform Representation of Enumerated Collections. Int J Funct Inform Personal Med 2010;3:273-91.

72. Mission & Principles. Clinical Data Interchange Standards Consortium, 2012. (Accessed November 15, 2012, at http://www.cdisc.org/mission-and-principles.)

73. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. J Am Med Inform Assoc 1996;3:224-33.

APPENDIX A

INSTRUMENT ELEMENT CODING MATRIX

Information Element	QS Terminology	LOINC	SNOMED CT	CognitiveAtlas	DOLCE COI	ETS	HaPI	MMY-TIP	PsycTESTS®	MeSH	Psychological Index Terms	PsycINFO®
Identifiers												
Instrument Name												
Instrument Acronym												
Instrument Alternative Name												
Instrument Author												
Instrument Author Type												
Instrument Adaptor												
Adapted From												
Instrument Release Date												
Instrument Primary Reference												
Physical Structure												
Instrument Language												
Instrument Readability Level												
Instrument Scale Name												
Instrument Scale Acronym												
Instrument Scale Count												
Instrument Subscale Name												
Instrument Item Length												
Item Name / Identifier												
Item Stimulus												
Item Response Format												
Item Response Stimulus												
Item Response Values												
Scientific Properties												
Respondent Relationship to Subject of												
Construct Measured												
Construct Qualifiers												
Construct Access Method												
Construct Target Time Frame												
Healthcare Domain												
Healthcare Objective												
Healthcare Activity												
Derived From												
Developmental Approach												
Developmental Model												
Evaluation Sample												
Psychometric Properties												
Validity												
Profile/Protocol Validity												
Construct Validity												
Content Validity												
Predictive Validity												
Convergent Validity												
Discriminative Validity												
Concurrent Validity												
Reliability												
Inter-rater Reliability												

Information Element	QS Terminology	LOINC	SNOMED CT	CognitiveAtlas	DOLCE COI	ETS	HaPI	MMY-TIP	PsycTESTS®	MeSH	Psychological Index Terms	PsycINFO®
Test-Retest Reliability												
Parallel-Forms Reliability												
Internal Consistency Coefficient												
Target Populations												
Respondent Population												
Respondent Age												
Respondent Gender												
Respondent Culture												
Respondent Setting												
Respondent Healthcare Status												
Respondent Social Role												
Respondent Exposures												
Scoring and Interpretation												
Instrument Scores Yielded												
Instrument Score Type												
Score Calculation												
Score Interpretation Method												
Instrument Factors												
Instrument Scoring Norms												
Instrument Logistics and Access												
Instrument Price												
Instrument Publisher Information												
Instrument Publisher Name												
Instrument Publisher Address												
Instrument Publication Status												
Instrument Access Source												
Instrument Languages Available												
Instrument Web Site												
Administration Time												
Administration Manual												
Associated Materials												
Administration Accommodations												
Administration Qualifications Required												
Instrument Use Permissions Required												
Instrument Development Funding Source												
Administration												
Administration Setting												
Administration Social Context												
Administration Medium												
Administration Method												
Item Delivery Method												
Administration Manual Used												
Instrument-Related Publications												
Publication Citation												
Publication Source												
Publication Type												
Publication Sample												

APPENDIX B

INSTRUMENT INFORMATION CLASSES AND ELEMENTS

Information Class	Information Element
Instrument Identifiers and Intrinsic Attributes	Instrument Name
	Instrument Acronym
	Instrument Alternative Name
	Instrument Author
	Instrument Author Type
	Adapted From
	Instrument Adaptor
	Instrument Release Date
	Instrument Primary Reference
Instrument Physical Structure	Instrument Scale Name
	Instrument Scale Acronym
	Instrument Scale Count
	Instrument Subscale Name
	Instrument Item Length
	Item Identifier
	Item Stimulus
	Item Response Stimulus
	Item Response Value
	Item Response Format
	Instrument Language
	Instrument Readability Level
Instrument Scoring and Interpretation	Instrument Scores Yielded
	Instrument Score Type
	Score Calculation
	Score Interpretation Method
	Instrument Factors
	Instrument Population Norms
Instrument Administration	Administration Setting
	Administration Social Context
	Administration Medium
	Administration Method
	Item Delivery Method
	Administration Manual Used
Instrument Target Populations	Respondent Population
	Respondent Age
	Respondent Gender
	Respondent Culture
	Respondent Setting
	Respondent Healthcare Status
	Respondent Social Role
	Respondent Exposures
Instrument Scientific Properties	Construct Measured
	Construct Qualifiers

Information Class	Information Element
	Construct Access Method
	Construct Target Time Frame
	Respondent Relationship to Subject of Test
	Healthcare Domain
	Healthcare Objective
	Healthcare Activity
	Derived From
	Developmental Approach
	Developmental Model
	Evaluation Sample
	Psychometric Properties
	Validity
	Profile/Protocol Validity
	Construct Validity
	Predictive Validity
	Convergent Validity
	Discriminative Validity
	Content Validity
	Construct Validity
	Concurrent Validity
	Reliability
	Inter-rater Reliability
	Test-Retest Reliability
	Parallel-Forms Reliability
	Internal Consistency Reliability
Instrument Access and Logistics	Instrument Price
	Instrument Publisher Information
	Instrument Publisher Name
	Instrument Publisher Address
	Instrument Publication Status
	Instrument Access Source
	Instrument Languages Available
	Instrument Web Site
	Instrument Administration Time
	Instrument Administration Manual
	Instrument Associated Materials
	Administration Accommodations
	Administration Qualifications Required
	Instrument Use Permissions Required
	Instrument Development Funding Source
Instrument Publications	Publication Citation
	Publication Source
	Publication Type
	Publication Sample

APPENDIX C

CONSTRUCT CATEGORIES FOR INSTRUMENT SAMPLE

Construct	Type	Source
Generalized Anxiety Disorder	Disorder	NIMH
Obsessive-Compulsive Disorder (OCD)	Disorder	NIMH
Panic Disorder	Disorder	NIMH
Post-Traumatic Stress Disorder (PTSD)	Disorder	NIMH
Social Phobia (Social Anxiety Disorder)	Disorder	NIMH
Generalized Anxiety Disorder	Disorder	NIMH
Attention Deficit Hyperactivity Disorder (ADHD, ADD)	Disorder	NIMH
Autism Spectrum Disorders (Pervasive Developmental Disorders)	Disorder	NIMH
Bipolar Disorder (Manic-Depressive Illness)	Disorder	NIMH
Borderline Personality Disorder	Disorder	NIMH
Depression	Disorder	NIMH
Eating Disorders	Disorder	NIMH
HIV & AIDS	Disorder	NIMH
Schizophrenia	Disorder	NIMH
Substance Use	Disorder	NIMH
Suicide Prevention	Disorder	NIMH
Personality / Self	Construct	Author Consensus
Impulsivity	Construct	Author Consensus
Emotion	Construct	Author Consensus
Extra Pyramidal Side Effects	Construct	Author Consensus

APPENDIX D

MASTER INSTRUMENT LIST

Instrument	Category
Abnormal Involuntary Movement Scale (AIM)	Extra Pyramidal Side Effects
Acquired Capability for Suicide Scale (ACSS)	Suicide
Addiction Severity Scale (ASI)	Substance Use
Agoraphobic Cognitions Questionnaire (ACQ)	Panic Disorder
Anxiety Severity Index (ASI)	Panic Disorder
Autism Diagnostic Interview – Revised (ADI-R)	Autism Spectrum Disorders
Autism Diagnostic Observation Schedule (ADOS)	Autism Spectrum Disorders
Barratt Impulsiveness Scale (BIS-11)	Impulsivity
Beck Anxiety Inventory (BAI)	Generalized Anxiety Disorder
Beck Depression Inventory II (BDI-II)	Depression
Beck Scale for Suicidal Ideation (BSS)	Suicide
Borderline Symptom List (BSL-23)	BPD
Brief Psychiatric Rating Scale (BPRS v 4.0)	Bipolar Disorder
Brown ADD Rating Scales	ADD / ADHD
Bulimia Test Revised (BULIT-R)	Eating Disorders
Buss-Durke Hostility Index (BDHI)	BPD
CAGE	Substance Use
Carroll Rating Scale for Depression	Depression
Child-Behavior Checklist (CBCL 4/16)	General
Clinician Administered PTSD Scale (CAPS)	PTSD
Columbia Suicide Severity Rating Scale (C-SSRS)	Suicide
Connors' Adult ADHD Rating Scales (CAARS)	ADD / ADHD
Davidson Trauma Scale (DTS)	PTSD
Difficulties in Emotion Regulation Scale (DERS)	Emotion
Dimensional Assessment of Personality Pathology - Basic Questionnaire (DAPP)	Personality / Self
Dimensional Yale-Brown Obsessive Compulsive Scale (DYBOCS)	OCD
Eating Disorders Inventory (EDI)	Eating Disorders
Eating Disorder Examination (EDE)	Eating Disorders
Edinburgh Postnatal Depression Scale (EPDS)	Depression
Hamilton Rating Scale for Anxiety (HAM-A)	Generalized Anxiety Disorder
Hamilton Rating Scale for Depression (HRSD)	Depression
HIV Dementia Scale (HDS)	HIV & AIDS
Liebowitz Social Anxiety Scale (LSSS)	Social Anxiety Disorder
Living with HIV	HIV & AIDS
McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD)	BPD
Michigan Assessment Screening Test Alcohol Drug (MAST-AD)	Substance Use
Mini International Neuropsychiatric Interview (MINI)	General
Mini-Mental State Examination (MMSE)	General
Minnesota Multiphasic Personality Inventory (MMPI-2)	Personality / Self
Modified HIV Dementia Scale (MHDS)	HIV & AIDS
Montgomery-Asberg Depression Rating Scale (MADRS)	Bipolar Disorder
Panic Disorder Severity Scale (PDSS)	Panic Disorder
Penn State Worry Questionnaire (PSWQ)	OCD
Pervasive Developmental Disorder Behavior Inventory (PDDBI)	Autism Spectrum Disorders
PHQ-9	Depression
Positive Affect - Negative Affect Scale (PANAS)	Emotion
Positive and Negative Syndrome Scale (PANSS)	Schizophrenia
Psychopathy Checklist - Revised (PCL-R)	Personality / Self
Psychotic Symptom Rating Scale (PSYRATS)	Schizophrenia
PTSD Checklist (PCL)	PTSD
Revised Diagnostic Interview for Borderline Personality Disorder (DIB-R)	BPD
Rorschach	Personality / Self
Rosenberg Self Esteem Scale (RSES)	Personality / Self
Schizophrenia Cognition Rating Scale (SCoRS)	Schizophrenia
Social Phobia and Anxiety Inventory (SPAI)	Social Anxiety Disorder
State-Trait Anxiety Inventory (STAI)	Generalized Anxiety Disorder
Symptom-Checklist-90-Revised (SCL-90-R)	General
Thematic Apperception Test (TAT)	Personality / Self

Instrument	Category
Toronto Alexithymia Scale - 20 (TAS-20)	Emotion
UPPS-P Impulsive Behavior Scale (UPPS-P)	Impulsivity
Wender Utah Rating Scale (WURS)	ADD / ADHD
Yale-Brown Obsessive Compulsive Scale (YBOCS)	OCD
Young Mania Rating Scale (YMRS)	Bipolar Disorder

APPENDIX E

INSTRUMENT SAMPLE

Instrument	Category
Abnormal Involuntary Movement Scale (AIM)	Extra Pyramidal Side Effects
Acquired Capability for Suicide Scale (ACSS)	Suicide
Addiction Severity Scale (ASI)	Substance Use
Agoraphobic Cognitions Questionnaire (ACQ)	Panic Disorder
Anxiety Severity Index (ASI)	Panic Disorder
Autism Diagnostic Interview – Revised (ADI-R)	Autism Spectrum Disorders
Autism Diagnostic Observation Schedule (ADOS)	Autism Spectrum Disorders
Barratt Impulsiveness Scale (BIS-11)	Impulsivity
Beck Anxiety Inventory (BAI)	Generalized Anxiety Disorder
Beck Depression Inventory II (BDI-II)	Depression
Beck Scale for Suicidal Ideation (BSS)	Suicide
Borderline Symptom List (BSL-23)	BPD
Brief Psychiatric Rating Scale (BPRS v 4.0)	Bipolar Disorder
Brown ADD Rating Scales	ADD / ADHD
Bulimia Test Revised (BULIT-R)	Eating Disorders
Buss-Durke Hostility Index (BDHI)	BPD
CAGE	Substance Use
Carroll Rating Scale for Depression	Depression
Child-Behavior Checklist (CBCL 4/16)	General
Clinician Administered PTSD Scale (CAPS)	PTSD
Columbia Suicide Severity Rating Scale (C-SSRS)	Suicide
Connors' Adult ADHD Rating Scales (CAARS)	ADD / ADHD
Davidson Trauma Scale (DTS)	PTSD
Difficulties in Emotion Regulation Scale (DERS)	Emotion
Revised Diagnostic Interview for Borderline Personality Disorder (DIB-R)	BPD
Rorschach	Personality / Self
Rosenberg Self Esteem Scale (RSES)	Personality / Self
Schizophrenia Cognition Rating Scale (SCoRS)	Schizophrenia
Social Phobia and Anxiety Inventory (SPAI)	Social Anxiety Disorder
State-Trait Anxiety Inventory (STAI)	Generalized Anxiety Disorder
Symptom-Checklist-90-Revised (SCL-90-R)	General
Thematic Apperception Test (TAT)	Personality / Self
Toronto Alexithymia Scale - 20 (TAS-20)	Emotion
UPPS-P Impulsive Behavior Scale (UPPS-P)	Impulsivity
Wender Utah Rating Scale (WURS)	ADD / ADHD
Yale-Brown Obsessive Compulsive Scale (YBOCS)	OCD
Young Mania Rating Scale (YMRS)	Bipolar Disorder

APPENDIX F

ATTRIBUTE CODING MANUAL FOR PSYCHOLOGICAL INSTRUMENT DATABASES, VERSION 1.0

Instrument Identifiers

Instrument identifiers are information elements that allow one to unambiguously identify an instrument. These include the *instrument name*, *instrument acronym*, and *alternative names*, as well as the *author*, *publication date* and *primary reference* for the publication in which the instrument was initially described. Instruments are often locally modified to meet the needs of a specific study or population[42]. When instruments are updated, revised or modified, the *instrument translator or adaptor* and the original instrument from which the current instrument was adapted are important identifying elements.

Many instruments are created by modifying and/or combining components of one or more existing instruments to create a new instrument. The element *derived from* is used to accommodate this information. In the current model, we distinguish between *simple modifications* (i.e., changes not intended to change the substantive properties of the instrument) and *adaptations* (i.e., changes that significantly alter the operational definition of the construct or constructs measured by the instrument). Both types of changes have a type (length, language, setting, construct, scale, response options, etc.) and a source instrument (i.e., the instrument to which, or from which, the modifications were made). However, simple modifications are modeled using the version element of the intrinsic properties schema, while adaptions are modeling using the derived from element of the scientific structure schema.

Instrument Name	The datasource provides the name of the instrument.
Instrument Acronym	The datasource provides the acronym (s) of the instrument.
Instrument Alternative Name	The datasource provides one or more alternative name(s) by which the instrument is known (in addition to the primary or formal name).
Instrument Author	The datasource provides a name for the author or developer of the test. If the instrument is a translation or adaption, it includes the name of the author of the original instrument from which the current instrument is adapted (and not just the name of the adaptor or translator).
Instrument Author Type	The datasource indicates whether the author provided is an individual, organization or institution.
Instrument Adapted or Translated From	If the instrument is a translation or adaption, the datasource includes the name of the instrument from which the current version was adapted or translated. Note: The current model distinguishes between two types of revisions or modifications made to instruments: adaptations and derivations.
Instrument Adaptor or Translator	The datasource provides the name of the individual, organization or institution that adapted or translated the current version.
Instrument Release Date	The datasource provides the date or year in which the instrument was published or released.
Instrument Primary Reference	The datasource provides the full reference (citation) for the primary source (i.e., the publication) in which the instrument was introduced, described and/or results of its evaluation reported. Note: There may be more than one primary source (as when different properties of the instrument are described in two different papers).
Instrument Development Funding Source	The datasource provides the name of the individual, institution or organization that funded development of the instrument

Instrument Physical Structure	
A number of elements in the model relate to the physical structure of instruments. These include *instrument scales*, *items* and *responses*. The *name* and *acronym* attributes have been added to both the instrument scale and instrument subscale entities. Instrument *short item name*, *item text*, and *item response text* are attributes of the instrument item entity. The *instrument readability level* is an intrinsic property of instruments, and is an important information element when it comes to instrument selection. In the current model, a *scale* is defined as a named or unnamed subset of one or more items within an instrument. Scales have a recursive relationship. That is, a scale can contain additional scales. In this model, the instrument itself is considered the container for all scales (and the entire set of all items is the top-level scale within the instrument). A scale is associated with one or more *healthcare constructs* and one or more *scores*. The *score* may be a quantitative metric or qualitative metric (e.g., the "dismissive attachment style" of the Adult Attachment Interview). A scale differs from a subscale, in that there is no single subset of items (scale) other than the instrument itself, to which all scale items belong. A subscale, on the other hand, consists of a subset of items within a single scale. An *item* is the most granular element of an instrument upon which a discrete evaluation can be made. In a questionnaire, an item is a single question. In a non-questionnaire type test, such as the *Rorschach Test*, an item is the smallest component of the instrument to which a *score* is assigned. Each item has an *item identifier* (typically an item number or name) and *item stimulus*. The item stimulus is the verbatim text for questionnaires; the image to which the respondent is asked to respond; or the link to the computer program which a respondent is asked to interact). Each item also has a number of item response attributes. The item response consists of the *response method* (e.g., Likert Scale, Q-sort, T/F, etc.) and the *response format*.	
Instrument Language	The datasource provides the name of the language in which the instrument is written. Note: Code as "N/A" for non-written instruments. Do not use information about the language in which the manual and/or scoring materials are available to code this field.
Instrument Readability Level	The datasource provides the reading level at which written components of the instrument are written. Note: Code as "N/A" for non-written instruments.
Instrument Scale Name	The datasource provides the name(s) of the primary scale(s) in the instrument. A scale is defined as a set of one or more items within the instrument.
Instrument Scale Acronym	The datasource provides the acronym corresponding to the primary scale(s) in the instrument
Instrument Scale Count	The datasource provides the number scales in the instrument
Instrument Subscale Name	The datasource enumerates the name(s) of instrument subscale(s) (i.e., scales within the top level scales)
Instrument Item Length	The datasource provides the number of items in the instrument; terms used to indicate the number of items, such as "short-form", "abbreviated version", or "long form" are considered *version qualifiers* and therefore are not coded here.
Item Identifier	The datasource provides a nominal (i.e., name) and/or numeric (i.e., item number) identifier for *each* item in the instrument. DO NOT use data included in the actual instrument to code this field for those datasources that provide a link to the actual instrument. This data *must* be included in one or more of the datasources' fields (i.e., one must be able to query the datasource for this information).
Item Stimulus	The datasource provides the stimulus used for each item in the instrument (e.g., the verbatim text for questionnaires; narratives or images to which the respondent is asked to respond; a link to the computer program with which a respondent is asked to interact). Do not use descriptions or sample items to code this field.
Item Response Format	The datasource provides the format in which the respondents' responds to items in the instrument ▪ Ratings - Binary (Y/N, T/F, Agree/Disagree), Likert Scale.

	▪ Item Selection ▪ Item Categorization or Classification (e.g., Q-Sort) ▪ Sentence Completion ▪ Creation of a Graphic ▪ Free-Form Narrative Response ▪ Visually-Assisted Rating Scale (NRS, VAS) DO NOT use data included in the actual instrument to code this field for those datasources that provide a link to the actual instrument. This data *must* be included in one or more of the datasources' fields (i.e., one must be able to query the datasource for this information). Do not use descriptions or sample items to code this field.
Item Response Stimulus	The datasource provides a full description (e.g., verbatim text for each of the possible responses in a multiple-choice questionnaire); Code as "N/A" for instruments for which item responses are not defined (e.g., observational measures, interviews, or sentence completion-type measures). Do not use descriptions or sample items to code this field.
Item Response Values	The datasource provides the value (or metric) corresponding to each *response stimulus* item (e.g., "never" = 0, "rarely" = 1, "sometimes" = 2, "often" = 3, "always" = 4; "yes" = 1, "no" = 2). Code as "N/A" for instruments for which item responses are not defined (e.g., observational measures, interviews, or sentence completion-type measures)

Instrument Scientific Properties	
The scientific properties of an instrument are first assessed during the development process and are continually assessed during the ongoing evaluation of the instrument's performance as it is used in research and clinical practice. The elements *developmental approach*, *developmental model*, and *construct access method* refer to the theoretical foundation of the test. The *developmental approach* refers to whether the scales and items in the instrument were developed empirically (e.g., blah, blah) or based on a theoretical conceptualization of the construct (e.g., blah, blah). The element *developmental model* refines and qualifiers the *developmental approach* element, by describing either the theoretical conceptualization upon which a theoretically-derived instrument is based, or describing the empirical method (e.g., factor analysis, multidimensional scaling) used to develop an empirically-derived instrument. Note: For constructs measured, think about HaPI description of PANSS ("Based on two established psychiatric rating systems, the PANSS was conceived as an operationalized, drug-sensitive instrument that provides balanced representation of positive and negative symptoms and gauges their relationship to one another and to global psychopathology")	
Respondent Relationship to Subject of Test	The datasource includes an information element describing the individual for whom the instrument is designed to be completed, and his or her relationship to the subject of the test (e.g., "parent version", "teacher form", "clinician-rated", "self-report")
Construct Measured	The datasource provides information about the construct(s) measured by the instrument (e.g., *depression*, *anxiety*, *suicidality*) A *behavioral health construct* is the class of behavioral or psychological entities measured by the instrument. A *behavioral health construct* may be an entire psychological "system" (such as personality, emotion, or cognition), a specific psychological construct, a sign or symptom, a disorder or syndrome, a level or area of functioning, or a diagnostic construct. Note: Constructs can have a parent-child relationship to other constructs. In other words, aspects, components, or dimensions of constructs are considered child constructs and not construct qualifiers
Construct Qualifier	The datasource provides detailed information about construct(s) measured by the instruments that includes qualifiers (e.g., depression *severity*, depression *onset*, anxiety *frequency*, *context of* suicide attempt; *differentiate between* depression and anxiety; etc.) A *construct qualifier* is used to further refine the *behavioral health construct*. Construct qualifiers include constructs related to the presence of symptoms or disorders (predisposition towards, differentiation/discrimination), the specific attributes of symptoms or disorders (e.g., severity, frequency, pattern, change, pattern); manifestation (contexts of occurrence, eliciting stimuli, and aggravating factors, ameliorating factors) or consequences (impact of functioning). Qualifiers that refine the construct (e.g., <u>motor</u> impulsivity, <u>non-planning</u> impulsivity, <u>catatonic</u> schizophrenia are not considered *construct qualifiers*, but rather constructs themselves) Note: Constructs can have a parent-child relationship to other constructs. In other words, aspects, components, or dimensions of constructs are considered child constructs and not construct qualifiers
Construct Access Method	The datasource provides information elements that describe the instrument's theoretical approach to accessing (or obtaining a measure of) the construct(s) measured by the instrument (e.g., "projective test", "semantic differential test", "inkblot test", "sentence completion method", "introspection"). <u>Note</u>: self-evaluation indicates introspection.
Construct Target Time Frame	The datasource provides information elements that describe the time frame, or period, for which the construct is measured (e.g., "current and lifetime", "retrospective", "in the prior 2 weeks", "6 months following the trauma", etc.)
Healthcare Domain	The datasource describes the healthcare domain in which the instrument is designed to

Instrument Scientific Properties	
	be used (e.g., epidemiological, clinical, research, community). This is the explicitly articulated domain of the instrument. Note: When coding dataset records, DO NOT assume that an instrument's domain is healthcare delivery because the purpose of the instrument relates to measuring a specific construct in patients only. The *healthcare domain* refers to specific domain or discipline for which the instrument is developed, or in which the instrument is designed to be used. These domains include psychometrics research, primary research, clinical research, healthcare delivery, health services research, epidemiological research, and quality measurement and reporting.
Healthcare Activity	The datasource describes the specific *healthcare activity* for which the instrument is designed to be used. The activity may be general (e.g., "psychological evaluation", "personality assessment", etc.) or specific (e.g., "determine prevalence of anxiety disorders", "assess the severity of depression", "assess suicide risk", "assess lethality of suicide attempts", "identify contexts in which panic attacks occur", etc.). A *healthcare activity* is any activity performed in one of the healthcare domains which involves the use of a psychological instrument. These activities include calculating the divergent or convergent validity of an instrument; determining prevalence and incidence of a specific symptom, disorder or syndrome in a community in epidemiological research; and assessing the existence, severity, duration or frequency of a specific symptom. Other examples of instrument-related healthcare activities include monitoring signs or symptoms, and collecting patient history data. Think of this attribute as answering the question "What are you doing with this measure?"
Healthcare Objective	The datasource describes the *healthcare objective* for which the instrument is designed to be used (e.g., to identify patient needs, to identify risk, to diagnose, to determine prognosis, to plan treatment, to determine effectiveness of intervention). The *healthcare objective* refers to the specific objective or goal the instrument is designed to meet. An objective always relates to a decision that can be made, or an action that can be taken on the basis of the information obtained through use of the measure. Objectives include making a diagnosis, or formulating a case. Other examples of objectives are determining a patients' prognosis, developing a treatment plan, monitoring response to treatment, and evaluating outcomes. Think of this attribute as answering the question "Why are using the measure? What is the goal to be achieved by using this measure?"
Derived From	The element *derived from* is used to describe the instrument from which the instrument was created. NOTE: Code superficial adaptations or modifications to instruments, such as length versions, language translations, or alternate forms (paper v interview) as an "adaptation" in the "instrument identifier" section. Code substantial modifications that change the psychometric properties of the instrument, or that result in essentially "new" instruments as derivations. (Essentially, if the older and newer versions are intended to measure the same construct in the same way. then code as "adaptation"; if the newer version is not comparable, b/c it represents a more valid or reliable measure of the original construct, then code as an "adaptation" b/c it is a designed to be a refinement and improvement of the original instrument; otherwise, code as "derivation). Example: The statement "The AGS version of the MSEL combines and updates the earlier Infant- and Preschool-MSEL versions" identifies the "Infant- and Preschool-MSEL versions" as the instrument from which it was derived.
Developmental Approach	The datasource provides information about the theoretical basis of the instrument (i.e., the approach guiding development of the instrument). The *developmental approach* refers to whether the scales and items in the instrument were developed empirically (e.g., based on factor analysis, or multi-dimensional scaling techniques) or based on a theoretical conceptualization of the construct (e.g., Erickson's theory of early

Instrument Scientific Properties	
	development, Freud's dream theory, Beck's cognitive theory, etc.). Example: The datasource might say that the instrument was based on a developmental theory, or social development theory (theory-based approach). Alternatively, it may say that an aggression scale was developed based on factor analysis of responses to questions provided to groups of convicted violent offenders and non-violent controls (empirically-based approach).
Developmental Model	The datasource provides information about the specific model, or method, used to develop the instrument. This information element refines, extends or qualifies the *developmental approach* element. Example: The datasource may say that the instrument was based specifically on "Erikson's theory of development" or "Piaget's theory" or "Freud's theory of repression as the source of neurosis". For empirically derived instruments, the datasource may indicate that "factor analysis" or "multi- dimensional scaling" was used to determine which items should be included in the instrument or a particular scale in the instrument.
Evaluation Sample	The datasource provides information about the specific population in which the instrument was evaluated prior to formal release or publication. NOTE: The evaluation sample is NOT the same as the target population in which the instrument is intended to be used (although these are likely identical or similar). The evaluation sample is the sample upon which the original psychometric properties of the instrument are calculated and reported.
Psychometric Properties Psychometric property elements are modeled at three levels of granularity to accommodate the way in which data is structured in instrument database fields. At the most general level a generic "psychometric properties" fields represents and stores whether - and which - psychometric properties have been evaluated, calculated or reported. The next level, validity and reliability, model and store data about whether - and which – validity and reliability properties have been evaluated, calculated or reported. The most granular level is the specific type of metric. This level includes *profile or protocol validity*, *construct validity*, *predictive validity*, *convergent validity*, *discriminative validity*, *construct validity*, *concurrent validity*, *inter-rater reliability*, *test-retest reliability*, *parallel-forms reliability*, and *internal consistency reliability*.	
Psychometric Properties	The datasource provides information about at least one psychometric property (i.e., some form of validity or reliability) of the instrument. NOTE: Some datasources have a field titled "psychometric properties" within which the detailed validity and reliability metrics are reported. In this case, we would code "psychometric properties" as being explicitly represented, and the detailed values as being implicitly represented. In cases where the datasource contains discrete fields for both validity and reliability, DO NOT CODE psychometric properties (leave it blank).
Validity	The datasource provides information about at least one measure of validity.
Reliability	The datasource provides information about reliability.
Profile / Protocol Validity	The datasource provides information about whether the instrument includes a check of the validity of the respondents' responses. Note: The existence of a "validity" scale or "faking" scale or item are examples of profile/protocol validity checks; Code any instrument that reports having one of these scales as "implicit"
Construct Validity	The datasource provides information about construct validity; construct validity consists of convergent and divergent validity. Construct validity is sometimes referred to as 'content validity'. Note: Code 'construct validity' in addition to 'convergent validity' and/or 'divergent validity' if either or both are provided in the datasource.
Convergent Validity	The datasource provides information about convergent validity. Convergent validity is a measure of the correlation between constructs that are expected to be related;

Instrument Scientific Properties	
(Construct Validity)	convergent validity is strengthened by finding a high correlation between the construct in question and constructs expected to be related to the construct in question. Note: Code 'construct validity' in addition to 'convergent validity' and/or 'divergent validity' if either or both are provided in the datasource.
Discriminative / Divergent Validity (Construct Validity)	The datasource provides information about discriminative validity. Discriminant validity (or divergent validity) is a measure of the correlation between constructs that are expected to have no relationship; divergent validity is strengthened by finding a low correlation between the construct in question and constructs expected to be unrelated to the construct in question. Note: Code 'construct validity' in addition to 'convergent validity' and/or 'divergent validity' if either or both are provided in the datasource.
Predictive Validity (Criterion Validity)	The datasource provides information about predictive validity. Predictive validity is a measure of the correlation between results obtained using the instrument with external variables (measured at a future point in time) with which they are expected to correlate.
Concurrent Validity (Criterion Validity)	The datasource provides information about concurrent validity. Concurrent validity is a measure of the correlation between results obtained using the instrument with concurrently measured external variables with which they are expected to correlate.
Content Validity	
Inter-Rater Reliability	The datasource provides information about inter-rater reliability. Inter-rater reliability is a measure of the correlation between results obtained using the instrument under the same conditions, at the same time, by two different raters.
Test-Retest Reliability	The datasource provides information about test-retest reliability. Test- retest reliability is a measure of the correlation between results obtained using the instrument under the same conditions, at two different points in time.
Parallel-Forms Reliability	The datasource provides information about parallel-forms reliability. Parallel-forms reliability is a measure of the correlation between results obtained using two different forms of the instrument (e.g., paper and computerized) under the same conditions, at the same time.
Internal Consistency Reliability	The datasource provides information about internal consistency reliability. Other terms commonly used to describe internal consistency reliability are "Cronbach's alpha" and "internal reliability".

Target Populations	
Note: One unresolved issue for future coding relates to groups. Groups like *elementary school students, middle school students, and high school students* seem to map well to the age group construct, although they also represent a social status or role; *college students*, and *graduate students* seem to map less clearly map to an age construct, and more clearly to a social status or role construct. Similarly, groups like *institutionalized individuals* and *prison inmates* are social roles or statuses as much as group settings; they could also arguably represent an exposure group]	
Respondent Population	The datasource provides information about the population *for which the instrument was developed*, or *in which it has come to be widely used.* This is different than the population in which the instrument is being used in a particular study (in which case, the population is the "evaluation sample").
Respondent Age	The datasource provides information about the age group for which the instrument was developed. This may be a specific age range (e.g., 4 – 8 year olds) or a generic age range (e.g., toddlers, school aged children, adolescents, older adults). [To think about: elementary school student, middle school students, high school students seem to fit with age, although also a social status or role; college students, and graduate students seem to map less clearly map to an age construct, and more clearly to a social status or role construct]
Respondent Gender	The datasource provides information about the gender for which the instrument was developed (e.g., males, females).
Respondent Culture	The datasource provides information about the culture or country for which the instrument was developed (e.g., American, European, etc.)
Respondent Setting	The datasource provides information about the setting (e.g., inpatient, community, school, prison, primary care) in which the instrument was developed to administered
Respondent Healthcare Status	The datasource provides information about the healthcare status of the population for which the instrument was developed (e.g., diagnostic status, symptom status, etc.)
Respondent Social Role	The datasource provides information about the social roles (e.g., parents, caregivers, prisoners, refugees) of the population for which the instrument was developed.
Respondent Exposures	The datasource provides information about the exposures (e.g., childhood trauma, military, combat) of the population for which the instrument was developed

Scoring and Interpretation	
Instrument Scores Yielded	The datasource enumerates the scores yielded by the test (e.g., a total score, global index, subscale scores, etc.)
Instrument Score Type	The datasource provides information about the type of score provided (e.g., raw score, profile score, aggregate score, t-score, index, etc.)
Score Calculation	The datasource provides information about how the score is calculated.
Score Interpretation Method	The datasource provides information about how the score is interpreted; information about how the score (or metric) translates into a clinically meaningful observation or finding
Instrument Factors Yielded	The datasource provides information about factors yielded by the test
Instrument Population Norms	The datasource provides information about population norms.

Instrument Logistics and Access	
Instrument Price	The datasource provides information about the price of the instrument.
Instrument Publisher Information	The datasource provides information about the publisher. Use this code if the datasource if all publisher information (name, address, etc.) is included in a single field
Instrument Publisher Name	The datasource provides the publisher name
Instrument Publisher Address	The datasource provides the publisher address
Instrument Publication Date	The datasource provides the instrument publication year
Instrument Publication Status	The datasource provides information about whether the instrument was ever published, whether it is currently in print and/or available.
Instrument Access Source	The datasource provides information about where or how to access the instrument (e.g., hyperlink, web site or mailing address).
Instrument Languages Available	The datasource provides information about the alternative languages in which the instrument is available.
Instrument Web Site	The datasource provides the web site at which the instrument can be accessed
Administration Time	The datasource provides information about how long it takes to administer the instrument.
Administration Manual	The datasource provides a name or citation for any administration manuals available for the instrument.
Associated Materials	The datasource provides information about materials that come with or can be ordered with the instrument
Administration Accommodations	The datasource provides information about the types of accommodations available (e.g., large print, etc.) available for administering the instrument
Administration Qualifications Required	The datasource provides information about what qualifications are required to administer the instrument
Instrument Use Permissions Required	The datasource provides information about what permissions are required to use or administer the instrument

Administration	
Unresolved: Typically interviews are "individual" (social context) and "face-to-face" (medium); for purposes of the current study, we will code these values only when explicitly stated in the record. Think about what attributes are modeled at what level – instrument, scale, item level; eg: MMSE, multi-modal (section 1, verbal query; section 2, "tests ability to name, follow verbal and written commands, write a sentence spontaneously, and copy a complex polygon"); additional attributes of item delivery method (e.g., PSWQ, some items presented in "reversed fashion")	
Administration Setting	The administration setting is defined as the setting in which the instrument is designed to be administered (e.g., classroom, workplace, natural environment, clinic, or hospital)
Administration Social Context	The administration social context is the social context in which the instrument is designed to be delivered, such as individual or group.
Administration Media Format	The administration medium is defined as the medium in which the instrument or scale is designed to be delivered, or in which it is created. Values include pen-and-paper (such as written questionnaires, surveys, checklists, etc.), face-to-face (such as during a face-to-face interview method), audio, audio video, etc. Examples/Issues ▪
Administration Method	The *administration method* is defined as the overall approach taken to obtain the measurement data. Values include querying the respondent (e.g., interview, survey, questionnaire,), observation, manipulation of objects, and interaction with media, environment or technology.
Item Delivery Method	The item delivery method is defined as the method by which individual items are delivered. Values include structured, semi-structured, skip-logic or unstructured.
Item Response Method	The item response method defines the specific manner in which the respondent is asked to provide his or her response. Values include true/false, Likert scale, semantic differential, sentence completion, or visual analog scale. Examples: "statements are rated on a 7-point scale"
Accommodations	Accommodations refer to specific modifications made to the procedure for specific respondent populations
Administration Manual Used	The datasource provides the name or a citation for the manual used in the evaluation of the instrument. The administration manual is a specific, written protocol used to describe a standardized administration process.

Instrument-Related Publications	
Publication Citation	The datasource provides citations for instrument reviews or publications
Publication Source	The datasource provides information about where it obtained information about the instrument (e.g., the publisher, the literature, a specific handbook or database, etc.)
Publication Type	The datasource provides information about the type of instrument publication (e.g., whether it is the original publication describing development of the instrument, a publication in which the instrument was used, a publication formally reviewing the instrument, etc.)
Publication Sample	The datasource provides information about the sample used in the publication
Publication References	The datasource provides a list of references associated with the publication