

University of Massachusetts Medical School

eScholarship@UMMS

---

GSBS Dissertations and Theses

Graduate School of Biomedical Sciences

---

2021-03-23


## Measurement in Health: Advancing Assessment of Delirium

Benjamin K.I. Helfand

*University of Massachusetts Medical School*

Let us know how access to this document benefits you.

Follow this and additional works at: [https://escholarship.umassmed.edu/gsbs\\_diss](https://escholarship.umassmed.edu/gsbs_diss)

 Part of the [Geriatrics Commons](#), [Nervous System Diseases Commons](#), [Pathological Conditions, Signs and Symptoms Commons](#), and the [Psychiatry and Psychology Commons](#)

---

### Repository Citation

Helfand BK. (2021). Measurement in Health: Advancing Assessment of Delirium. GSBS Dissertations and Theses. <https://doi.org/10.13028/vtns-j677>. Retrieved from [https://escholarship.umassmed.edu/gsbs\\_diss/1122](https://escholarship.umassmed.edu/gsbs_diss/1122)

Creative Commons License



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 License](#)

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in GSBS Dissertations and Theses by an authorized administrator of eScholarship@UMMS. For more information, please contact [Lisa.Palmer@umassmed.edu](mailto:Lisa.Palmer@umassmed.edu).

MEASUREMENT IN HEALTH: ADVANCING ASSESSMENT OF DELIRIUM

A Dissertation Presented

By

BENJAMIN KEVIN INOUYE HELFAND

Submitted to the Faculty of the  
University of Massachusetts Graduate School of Biomedical Sciences, Worcester  
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

MARCH 23, 2021

MD/PhD PROGRAM

MEASUREMENT IN HEALTH: ADVANCING ASSESSMENT OF DELIRIUM

A Dissertation Presented  
By

BENJAMIN KEVIN INOUYE HELFAND

This work was undertaken in the Graduate School of Biomedical Sciences  
MD/PhD Program  
Under the mentorship of

Richard N. Jones, Sc.D.; Thesis Advisor

Edwin D. Boudreaux, Ph.D.; Thesis Advisor

David D. McManus, MD.; Member of Committee

Chad E. Darling, MD.; Member of Committee

Bruce A. Barton, Ph.D.; Member of Committee

Alden L. Gross, Ph.D.; External Member of Committee

David A. Smelson, Psy.D.; Chair of Committee

Mary Ellen Lane, Ph.D.,  
Dean of the Graduate School of Biomedical Sciences

March 23, 2021

**DEDICATION**

This work is dedicated to the memory of:

**JOSHUA BRYAN INOUE HELFAND,  
DANIEL STEVEN SNYDER,  
STEVEN FRANCIS HAMILTON,  
RITA RACHEL ZELDA WOLFF HELFAND,  
MITSUO INOUE,  
ICHIRO INOUE,  
BRADLEY YOSHIO INOUE,  
&  
JON MASAMITSU INOUE**

## ACKNOWLEDGEMENTS

This entire project would not have been possible without the unwavering support of my mentors, Dr. Richard N. Jones, ScD and Dr. Edwin D. Boudreaux, PhD. Rich, I can't believe we've been working together for so long now. You've taught me so much, not only in terms of content of things like measurement, but in being a better researcher. I've learned so much along this path and it is thanks to your guidance. Dr. Boudreaux, I really appreciate how you have stuck by and supported me. You are a great inspiration in being a successful researcher.

I am grateful for the continued guidance and feedback of the members of my Thesis Research Advisory Committee (TRAC): Dr. David Smelson, who has served as my chair for both my TRAC and Dissertation Examination Committee (DEC) and has constantly supported my endeavors and has been a great sounding board; and Dr. David McManus, a true role model as an academic physician scientist. I am also so thankful to Drs. Bruce Barton, Chad Darling, and Alden Gross from Johns Hopkins University for serving on my DEC.

Thank you so very much to Provost Terrence Flotte and Dean Sonia Chimienti. I cannot even begin to express how thankful I am to have met such caring, thoughtful, and helpful leaders. I am so grateful that you have helped and encouraged me along my path.

Thank you to my clinical mentor, Dr. Gary Blanchard, I have really valued our conversations and your help and words of wisdom have been extremely valuable.

Thank you to my family, friends, and all my mentors. Your constant support and love have helped propel me to reach my goals. I never could have accomplished the things I have done in my life without my great support network.

Thank you so much to my parents and brother. I never could have done this without your constant love and support in all my endeavors. I've appreciated all the help in working towards this goal and the time you have devoted to helping me in any way along this journey. Thank you for the edits, I know red is love.

Finally, thank you to my wife, Danielle. What a crazy year it has been. There was no one I would rather have been together with nonstop than yourself, Simon, and Tuukka. Your love and support helps me overcome anything and reach for the stars. I cannot wait for us to grow our family and continue our journey together.

## ABSTRACT

*Rationale:* Delirium is a serious, morbid condition affecting 2.6 million older Americans annually. A major problem plaguing delirium research is difficulty in identification, given a plethora of existing tools. The lack of consensus on key features and approaches has stymied progress in delirium research. The goal of this project was to use advanced measurement methods to improve delirium's identification.

*Aims and Findings:*

- (1) Determine the 4 most commonly used and well-validated instruments for delirium identification. Through a rigorous systematic review, I identified the Confusion Assessment Method (CAM), Delirium Observation Screening Scale (DOSS), Delirium Rating Scale-Revised-98 (DRS-R-98), and Memorial Delirium Assessment Scale (MDAS).
- (2) Harmonize the 4 instruments to generate a delirium item bank (DEL-IB), a dataset containing items and estimates of their population level parameters. In a secondary analysis of 3 datasets, I equated instruments on a common metric and created crosswalks.
- (3) Explore applications of the harmonized item bank through several approaches. First, identifying different cut-points that will optimize: (a) balanced high accuracy (Youden's J-Statistic), (b) screening (sensitivity), and

(c) confirmation of diagnosis (specificity) in identification of delirium. Second, comparing performance characteristics of example forms developed from the DEL-IB.

*Impact:* The knowledge gained includes harmonization of 4 instruments for identification of delirium, with crosswalks on a common metric. This will pave the way for combining studies, such as meta-analyses of new treatments, essential for developing guidelines and advancing clinical care. Additionally, the DEL-IB will facilitate creating big datasets, such as for omics studies to advance pathophysiologic understanding of delirium.



**TABLE OF CONTENTS**

<b>Dedication</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>Abstract</b>	<b>vi</b>
<b>Table of Contents</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>x</b>
<b>List of Abbreviations</b>	<b>xi</b>
<b>List of Copyrighted Materials</b>	<b>xiii</b>
<b>Chapter I – Introduction</b>	<b>1</b>
<b>Chapter II – Detecting Delirium: A systematic review of identification instruments for non-ICU settings</b>	<b>23</b>
<b>Chapter III – Harmonization of four delirium instruments: Creating crosswalks and the Delirium Item-Bank (DEL-IB)</b>	<b>82</b>
<b>Chapter IV – Delirium Item Bank (DEL-IB): Utilization to Evaluate and Create Delirium Instruments</b>	<b>118</b>
<b>Chapter V – Discussion &amp; Future Directions</b>	<b>147</b>
<b>Bibliography</b>	<b>162</b>

## LIST OF TABLES

Table 1.1	Delirium diagnosis: DSM-5 definition
Table 1.2	Stevens' classification of measurement
Table 1.3	Performance characteristics of measurement instruments
Table 1.4	Definitions of common IRT terms
Table 2.1	Characteristics of articles reviewed
Table 2.2	Selection criteria for delirium identification instruments based on the original citation
Table 2.3	Comparison of 4 recommended delirium instruments (alphabetical order)
Table 2.4	COSMIN score of delirium identification instruments
Table 2.5	Comparison of delirium instruments (alphabetical order)
Table 2.6	Search strategies for databases
Table 2.7	COSMIN-guided psychometric review
Table 2.8	List of citations of eligible articles
Table 2.9	List of excluded instruments with reasons
Table 2.10	List of citations of all instruments included
Table 3.1	Baseline characteristics of the three datasets
Table 3.2	Kappa statistics of delirium identification between CAM (short), DOSS, DRS-R-98, MDAS
Table 3.3	DOSS crosswalk
Table 3.4	CAM Short-form crosswalk
Table 3.5	CAM Long-form crosswalk
Table 3.6	DRS-R-98 Severity crosswalk
Table 3.7	DRS-R-98 Total crosswalk
Table 3.8	MDAS crosswalk
Table 4.1	Baseline characteristics of the three datasets
Table 4.2	Instrument cut-points
Table 4.3	Four example instruments from the DEL-IB (Delirium Item Bank), each ordered by highest information
Table 4.4	Psychometric properties of proposed new instruments

## LIST OF FIGURES

- Figure 2.1** Systematic review flow diagram
- Figure 2.2** Domain coverage of 4 recommended delirium instruments
- Figure 3.1** Data structure and models
- Figure 3.2** Expected score characteristic curves of each delirium identification instrument
- Figure 3.3** Reliability of each delirium identification instrument
- Figure 4.1** ROC curves for each delirium identification instrument compared to DSM-5 criteria
- Figure 4.2** ROC curves for each example instrument compared to DSM-5 criteria

## LIST OF ABBREVIATIONS

<b>Abbreviation</b>	<b>Meaning</b>
3D-CAM	3-Minute Diagnostic Assessment
AUC	Area under the curve
BASIL	Better Assessment of Illness study
CAT	Computerized adaptive testing
CAM	Confusion Assessment Method
CAM-ICU	Confusion Assessment Method for the Intensive Care Unit
CHART-DEL	Chart Delirium Identification
CINAHL	Cumulative Index to Nursing and Allied Health Literature
COSMIN	Consensus-based Standards for the Selection of Health Measurement Instruments
DEL-IB	Delirium Item Bank
DOSS	Delirium Observation Screening Scale
DRS	Delirium Rating Scale
DRS-R-98	Delirium Rating Scale-Revised-98
DSM-III	Diagnostic and Statistical Manual, Third Edition
DSM-III-R	Diagnostic and Statistical Manual, Third Edition, Revised
DSM-IV	Diagnostic and Statistical Manual, Fourth Edition
DSM-IV-TR	Diagnostic and Statistical Manual, Fourth Edition, Text Revision
DSM-5	Diagnostic and Statistical Manual, Fifth Edition

EMBASE	Excerpta Medica Database
FAM-CAM	Family-Confusion Assessment Method
ICC	Intra-class correlation coefficient
ICD-10	International Classification of Diseases
ICU	Intensive care unit
IOM	Institute of Medicine
IQ	Intelligence quotient
IRT	Item Response Theory
IV	Intravenous
MCAT	Medical College Admission Test
MDAS	Memorial Delirium Assessment Scale
NEECHAM	Neelon and Champagne Confusion Scale
NIH	National Institute of Health
NPV	Negative Predictive Value
Nu-DESC	Nursing Delirium Screening Scale
PPV	Positive Predictive Value
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
PROMIS	Patient-Reported Outcomes Measurement Information System
ROC	Receiver Operating Characteristic Curve
SEM	Standard Error of Measurement

## List of Copyrighted Materials Produced by the Author

Table 1.1 is reprinted with permission from the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, (Copyright 2013). American Psychiatric Association.

Chapter II is adapted from a published manuscript with permission:

**Helfand, B. K.**, D'Aquila, M. L., Tabloski, P., Erickson, K., Yue, J., Fong, T. G., Hshieh T. T, Metzger, E. D., Schmitt, E. M., Boudreaux, E. D., Inouye, S. K. & Jones, R. N. (2021). Detecting Delirium: A Systematic Review of Identification Instruments for Non-ICU Settings. *Journal of the American Geriatrics Society*, 69(2), 547-555.

Chapter III is adapted from a manuscript in preparation and is included with permission not required.

Chapter IV is adapted from a manuscript in preparation and is included with permission not required.

## CHAPTER I – Introduction

### ***Why Measurement is Important***

A key facet of evidence-based medicine is the generation of new scientific evidence. The derivation of this evidence originates from direct observations and empirical research studies. A key component of these empirical research studies is translating clinical observations into data that one can use to compare groups, evaluate treatments, or elucidate clinical outcomes. Generally, this numerical data requires the use of measurement instruments. The science behind measurement informs the optimal construction and character of measurement instruments for specific purposes and in specific contexts. For example, even before testing in a study population, careful thought must be given to the instrument design in terms of who would administer the instrument (e.g., physicians, nurses, trained researchers), in what clinical context (e.g., medical or surgical wards, emergency department, intensive care unit (ICU)), and for what purpose (e.g., screening, diagnosing, severity rating). Optimizing measurement is important because it maximizes the accuracy of instruments and their efficient application, that is, allowing for the best use of limited clinical resources in the most efficient way.

Measurement is foundational to the field of delirium. Delirium, an acute change in cognition, characterized by a waxing and waning course with multiple cognitive impairments including inattention and disorientation, is still a clinical diagnosis without known laboratory tests or biomarkers. Thus, the measurement of delirium must inform all clinical and research developments in the field to assure decision-making is accurate and evidence-based. The field of delirium provides useful examples of how measurement must progress and adapt over time. The Confusion Assessment Method (CAM), developed in 1990, was one of the first measurements of delirium diagnosis (1). As the field has progressed, new and adapted measures have developed to serve new uses. For example, the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) (2, 3) was developed to measure delirium in the ICU setting and the CAM-S (4) was devised to accurately measure delirium severity.

### ***Clinical Overview of Delirium***

Delirium is an important yet under-recognized syndrome characterized by acute onset, inattention along with other cognitive impairments, and a waxing and waning course. Accounts of delirium date back several millennia (5) and was first used as a medical term in the first century AD, characterizing mental syndromes following head trauma or fever (6, 7). Since that time, several terms have emerged to describe delirium including: acute brain dysfunction, acute brain failure, acute brain



syndrome, acute cerebral insufficiency, acute confusion, acute confusional state, acute organic brain syndrome, acute organic psycho-syndrome, acute psycho-organic syndrome, clouding of consciousness, clouded state, metabolic encephalopathy, and toxic–metabolic encephalopathy (6, 8, 9). Each of these concepts influenced the ultimate definition of delirium that was codified in the Diagnostic and Statistical Manual (DSM-III) of the American Psychiatric Association in 1974, and regularly updated to the current version in DSM-5 (8, 10, 11). The full definition is shown in **Table 1.1**.

**Table 1.1. Delirium diagnosis: DSM-5 definition**

A. A disturbance in attention (i.e., reduced ability to direct, focus, sustain, and shift attention) and awareness (reduced orientation to the environment).
B. The disturbance develops over a short period of time (usually hours to a few days), represents a change from baseline attention and awareness, and tends to fluctuate in severity during the course of a day.
C. An additional disturbance in cognition (e.g., memory deficit, disorientation, language, visuospatial ability, or perception).
D. The disturbances in Criteria A and C are not better explained by another preexisting, established, or evolving neurocognitive disorder and do not occur in the context of a severely reduced level of arousal, such as coma.
E. There is evidence from the history, physical examination, or laboratory findings that the disturbance is a direct physiological consequence of another medical condition, substance intoxication or withdrawal (i.e., due to a drug of abuse or to a medication), or exposure to a toxin, or is due to multiple etiologies.

Reprinted with permission from the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, (Copyright 2013). American Psychiatric Association.

Delirium has far-reaching clinical and public health importance, yet it is an understudied neuropsychiatric disorder, especially when considering its overall impact on patients and healthcare systems (12). Delirium most commonly affects

individuals 65 and older, affecting over 2.6 million older Americans every year (12). It causes major burden and distress to patients, their caregivers and families, and healthcare professionals (13). Specifically, delirium is associated with increased length of hospital stay, increased rates of admission to long-term care institutions, and increased subsequent risk of developing dementia (12, 14). Mortality rates reach one quarter to one third of patients within two years of an episode of delirium (15). Medicare expenditures approach over \$160 billion annually for excess healthcare expenditures attributable to delirium in the United States (16). In the year following an episode, the average delirious patient costs over \$60,000 more to the healthcare system than those who did not develop delirium after adjustment for relevant confounders (17). Importantly, delirium is preventable in many cases (18).

Despite its clinical importance, recognition of delirium remains a major problem. A recent study showed 61% of hospitalized patients confirmed to have delirium by a palliative care expert, had the delirium diagnosis missed by the primary referring team (19). At least part of the problem with recognition of delirium has been attributed to the lack of a unified instrument for its identification. Thus, the development of a widely accepted, unified identification instrument for delirium would greatly assist with recognition, and would help with prevention and management of this common, morbid, and often fatal condition.

Delirium has risen in importance for public health during the 2020 global pandemic of COVID-19. A recent study found over a quarter of older persons presenting to the emergency department with COVID-19 infection have delirium, which was the sixth most common presenting symptom overall (20). Another study found delirium was present in a third of hospitalized patients with COVID-19 (21). It is imperative to recognize that delirium is a common, atypical presentation of COVID-19 or other viral infections. The consistent and accurate identification of delirium is critical and depends on valid measurement instruments. **Thus, delirium is clearly an important medical condition with far-reaching public health implications, and this project focused on applying advanced measurement methods to improve the identification of delirium.**

### ***Overview of Measurement***

Measurement has played a vital role in human survival and the flourishing of civilization since prehistoric times. For example, at the origin of agrarian society, measurement was necessary to determine how many crops to grow to feed the community (22). Many of the original units of measurement were based on natural biological objects (i.e., fingers to represent length), however, such measurements tended to vary according to local circumstances and needs. Variability can be dealt with by taking the average of many biological objects or consistent use of more standard and unvarying physical objects. Another major problem in the early

application of measurement was a lack of consistency in units. Historically, it was found that towns within the same country would use units of measurement that were the same in name, but the actual quantity was unstandardized and differed between towns. Many countries used different units making trade and exchange across countries difficult. Thus, a unified system, known as the *Système International d'Units* or SI units, was ultimately developed (22).

### ***Classification of Measurement Approaches***

In 1946, the psychologist S.S. Stevens provided a fundamental advance in the field of measurement by classifying and defining different types of measurement approaches that could be applied widely (23). He defined measurement “as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement” (23). These different kinds of scales included the following categories of measurement: nominal, ordinal, interval, and ratio (23) (**Table 1.2**). *Nominal measures* involve no inherent order in the categories. *Ordinal measures* do have an order in their categorization, but there is not a specific or consistent difference between each category. In *interval measures*, both the order and difference are specified and consistent between categories. *Ratio measures* are specific extensions of interval measures, where the value zero specifies the complete absence of what is being measured—an absolute zero. The end result

of these categorizations has been the development of mathematical derivations for each type of measurement, such that different categories or levels of the scale can be quantified numerically.

In the present work, I used item response theory (IRT), which places a latent trait estimate on an interval scale (more background on IRT is provided below). This allows for the direct comparison of differences both between participants and scores simultaneously, which cannot occur with an ordinal scale. An example of a nominal measure is the Confusion Assessment Method (CAM) that uses an algorithm in the case identification of delirium versus no delirium (1). An example of an ordinal measure is any scale that uses a sum score, such as the Memorial Delirium Assessment Scale (MDAS) (24).

**Table 1.2. Stevens' classification of measures**

Measurement type	Definition	Example(s)
Nominal	No inherent order in the categories	Hair color; Algorithm of case identification of delirium versus no delirium
Ordinal	Categories ordered, but no specific or consistent difference between each category	Wong-Baker Faces Pain Rating Scale; Likert scale; Sum scores on delirium identification instruments
Interval	Both the order and difference are specified and consistent between categories	Intelligence quotient (IQ); Temperature in Fahrenheit; IRT based estimates of the latent trait, propensity to delirium
Ratio	Specific kind of interval measures with absolute zero	Age; Height; Weight; Temperature in Kelvin

Definitions adapted from (22, 23).

### ***Clinical Measurement: Understanding Performance Characteristics***

This dissertation will focus on the importance of measurement to develop clinically useful measures to identify disease states. This section will elucidate topics and terms critical to evaluating the performance of a measurement instrument. A *construct* is an idea that contains key conceptual elements; in medicine, a construct is typically comprised of the signs and symptoms of a specific disease or disorder. In developing and evaluating measurement instruments, there are two key performance characteristics to understand: *reliability* and *validity*. Reliability of a measure refers to the consistency in findings on repeated measurements when

the patients have not changed (25, 26). Validity describes if a measurement instrument truly measures the construct it intends to measure (25, 26). Several different types of both reliability and validity exist, as detailed in **Table 1.3**.

**Table 1.3. Performance characteristics of measurement instruments**

Test characteristic	Description (or definition)	How assessed
Reliability	Consistency in findings on repeated measurements when the patients have not changed	Minimal measurement error – intra-class correlation coefficient (ICC), standard error of measurement (SEM), Cohen’s kappa, and McDonald’s Omega
Internal consistency reliability	Inter-relatedness between different sets of items in a measurement instrument	Standard error of measurement (SEM), McDonald’s Omega
Test-retest reliability	Comparing scores on an instrument over time with repeated testing, contributions from inter-rater and intra-rater reliability	Intra-class correlation coefficient (ICC)
Inter-rater reliability	Similarity of ratings of different observers making observations of the same patient at the same time	Intra-class correlation coefficient (ICC), Cohen’s kappa
Intra-rater reliability	Similarity of ratings the same observer at different time points	Intra-class correlation coefficient (ICC), Cohen’s kappa
Validity	If a measurement instrument truly measures the construct it intends to measure	See specific validity example types below
Content validity	Whether the subject matter and specific questions in an instrument correspond with the intended construct; both in terms of the relevance to the construct and measuring the full scope of the construct	Subjective assessment of the extent to which the instrument contains relevant items that assess domains of the construct

Face validity	Part of content validity that shows that an instrument properly reflects the planned construct, generally determined by experts in the field	Subjective assessment on the part of test users that the test is measuring the full construct
Criterion validity	How well the instrument compares to the reference standard	Correlation coefficients, sensitivity, specificity, predictive value
Predictive validity	Prediction of expected clinical outcomes	Mean differences, correlation coefficients, risk ratios, odds ratios
Convergent validity	Scores that agree with measures on existing tests of the same construct	Correlation coefficients, measures of agreement
Construct validity	Extent to which an instrument adequately measures the idea or concept of interest	Totality of the evidence for reliability and validity

Definitions adapted from (25, 26)

*Internal consistency reliability* refers to the inter-relatedness between different sets of items in a measurement instrument (27). *Inter-rater reliability* refers to the similarity of ratings of different observers making observations of the same patient at the same time, while *intra-rater reliability* shows the similarity of ratings the same observer at different time points (27). When scores on an instrument are compared over time with repeated testing, that is known as *test-retest reliability* (27). To know a measure is reliable, it must have minimal measurement error, which can be calculated with statistics such as intra-class correlation coefficient (ICC), standard error of measurement (SEM), Cohen's kappa, and McDonald's Omega (27).

*Content validity* pertains to the subject matter of the specific questions and items within the instrument, and whether they correspond with the intended construct to



be measured. Content validity covers both the relevance to the construct and measuring the full scope of the construct (for example administering items about all signs and symptoms relevant to a syndrome) (27). *Face validity* is a part of content validity that shows that an instrument properly reflects the planned measured disease (27), generally determined by experts in the field. In cases when the construct has a reference standard measurement, *criterion validity* refers to how well the new instrument compares to the reference standard in identifying the disease (27). *Construct validity*, originally coined for use in psychological tests by Cronbach and Meehl (28), is considered the fundamental aspect of validity and broadly means the extent to which an instrument adequately measures the idea or concept of interest (25, 29, 30). Construct validity can be informed by prediction of expected clinical outcomes (*predictive validity*) or scores that agree with measures on existing tests of same/similar constructs (*convergent validity*) (27).

### ***Clinical Measurement: Understanding Goals of Testing***

When constructing a test, it is important to know the intended usage of the test, and hence, design the test in such a way that will optimize that specific use. In other words, alternative uses of tests may lead the test designer to prioritize selecting different items during the construction of a new test for screening versus one for diagnosing or assessing severity. One should also put thought into if the desire is for inferences at the individual patient level or group-based. It is not

practical to develop a single test to serve all uses and all audiences simultaneously. In medical tests, these competing interests are often divided into discriminative and evaluative tests (31). In discriminative tests, one wants to distinguish differences between patients at a single point in time. In evaluative tests, one wants to measure the magnitude of construct, which can be useful in characterizing change over time. Therefore, the answers to questions on an evaluative test should change when health status changes, especially in relation to specific interventions or clinical events (32). More simply, discriminative tests define who is a case versus who is not a case, while evaluative tests show differences or change in health status (between groups, over time, or in response to treatment). It is possible to develop a test that can perform both these tasks; however, if the ultimate desire is only to address discriminative goals (e.g., diagnosis), then resources would be used inefficiently to also satisfy measurement properties that serve evaluative goals. For instance, participants would be asked too many questions, putting great burden on both the patients and the person administering the test. The test will not be adopted in practice if it is too long and burdensome to administer. A test needs to be as long as needed to do its job and no longer or shorter. In the present work, the interest was in delirium case identification, and thus, the ultimate goal was to construct discriminative tests.

***Clinical Measurement: Identification of Disease***

In clinical medicine, measurement plays a critical role in the accurate screening and diagnosis of disease states. It is important to understand the similarities and differences of the performance characteristics needed for diagnosis and screening tasks, and their implications for medicine and public health. In a clinical medicine context, screening occurs before the onset of disease and typically refers to tests performed in an asymptomatic or preclinical patient to help *prevent* the later onset of disease. Screening can also refer to asking a few, quick broad questions or ordering a few generic tests to determine a patient's level of risk for the disease. This can help the clinician to determine if more detailed questioning or testing is needed, such as identifying high-risk patients, who necessitate more close following and in which one might conduct early diagnostic testing. Diagnosis occurs after disease has occurred, and refers to the process of *identification* of the disease that explains the patient's signs or symptoms.

In a public health context, screening is often applied across a large group of people (i.e., at the population level) to assess which persons are at risk for a condition or already unknowingly have the condition, with a goal of helping to prevent the condition or its associated complications (33). For optimal screening, measurement instruments should select for maximal sensitivity, in order to avoid false negatives, or missing cases. Diagnosis refers to more detailed testing at an individual level, meant to confirm the existence of a condition in suspected patients

and to identify patients who need treatment. Diagnostic tests should select for maximal specificity, in order to avoid false positives and conducting more detailed or risky evaluations on cases without the condition.

However, there are important distinctions to consider for delirium diagnosis in a research context. Due to feasibility constraints, large research studies cannot typically use reference standard diagnoses, such as by trained physicians. Instead, the use of epidemiologic criteria to approximate reference standard diagnoses is often implemented in the research setting. One method utilized are expert panels to help define approximate reference standard approaches for diagnostic use in research studies (34). Expert panels can assist in deciding on rules to classify people according to likely diagnosis, incorporating the results of the research assessments and taking clinical judgment into account. This method has been shown to be superior to individual diagnoses and the best method when true reference standards (such as laboratory tests or histopathology) are not available (35). However, studies that use expert panels do not always have standardized approaches of defining diagnoses or for reporting their results, which limits their replication in other studies. Thus, standards need to be developed for how to adequately report a study using an expert panel.

In the context of delirium, there are additional challenges to screening and diagnosis. The formal diagnosis of delirium requires a detailed history, physical

examination, and laboratory testing, along with interviewing a knowledgeable informant who knows the patient well enough to report whether the current symptoms are a change from baseline. Currently, little is understood about the pathophysiology underlying delirium and no laboratory tests or imaging modalities exist as reference standards for diagnosis. The ability to fully assess the criteria laid out by the current reference standard, the DSM-5, requires an extensive time commitment and clinical expertise. Hence, this has led to the creation of many instruments to aid in the screening or diagnosis of delirium for a wide array of practitioners across clinical settings.

Screening for delirium has inherent challenges. There are no rigorous guidelines about who to screen and at what time point. Due to a lack of a true reference standard delirium diagnosis, there is no current way to compare performance of different screening instruments (16). As with any other form of screening, delirium screening should ideally help identify patients who do not have delirium, but are at risk for delirium, in order to prevent complications; however, often the delirium screening instruments are utilized inappropriately as diagnostic instruments. A plethora of instruments for screening across different clinical settings and practitioners have developed over time. One example is the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU), created to screen non-verbal (intubated) patients in the intensive care unit (ICU), since the rates of delirium are quite high in that clinical setting (2, 3). Similar screening instruments

have been created for use in the emergency department and in long-term care facilities (36, 37). Another example of a screening instrument is the Delirium Observation Screening Scale (DOSS), created to provide a quick and easy approach for nurses to screen their patients for delirium, without requiring a physician (38). A unifying characteristic of delirium screening instruments is that they must be quick and easy to use. Some widely used instruments can be used for either screening (CAM short form) or diagnosis (CAM long form) (1).

A major challenge in delirium is that the same instruments are used interchangeably for either screening and/or diagnosis, such that evaluation of studies can only focus on the combined “identification” of delirium. Thus, *identification of delirium* was selected as the focus of Chapter II, with a systematic review to comprehensively identify all delirium identification measures in active use. This first step to systematically inventory and evaluate all instruments in current use for identification of delirium was essential to select the best instruments for the subsequent analytic work.

### ***Item Response Theory (IRT)***

Item response theory (IRT) is an approach commonly utilized for advanced measurement development in healthcare, which evolved from work in educational testing in the 1950s-60s (39-41). Its application has greatly enhanced modern

psychometric research, and provides a robust approach to design and score measures (42). IRT defines a large grouping of statistical procedures created to associate discrete observations—such as responses to a questionnaire or symptom rating scale—to the underlying, but not directly observable *latent trait* (construct) presumed to cause the symptom. The key innovation of IRT is that it summarizes participants' levels on underlying traits and *test items* separately along the same scale (43). The focus of IRT is to estimate the latent trait (27), which one can consider as the tendency of participants to endorse or exhibit a symptom of disease.

While IRT is a powerful methodology for advancing measurement of constructs like delirium, it is important to understand some of its underlying assumptions. IRT assumes that a latent trait describes the *probability of correctly responding to an item or endorsing a symptom*, and that persons with a higher level on the latent trait have a higher probability of endorsing the symptom or answering the item correctly. A further assumption of *conditional independence* is fundamental to IRT, that is, the probability of a correct response is independent of other answers to items within the instrument, conditional on the level of the underlying latent trait (43). Another assumption involves the statistical and conceptual division of characteristics of symptoms (test items) and characteristics of participants. In IRT analyses, the latent trait is customarily assumed to have a normal distribution with

a mean of zero and standard deviation of one (44). Some key terms used in IRT analysis are presented and defined in **Table 1.4**.

**Table 1.4. Definitions of common IRT terms**

Item Response Theory (IRT) term	Description (or definition)
Latent trait	Tendency of participants to endorse a symptom
Conditional independence	Probability of a correct response is independent of other answers to items within the instrument, and is conditional on the level of the underlying latent trait
Difficulty parameter	Level of the trait at which a participant picked randomly from the study population has a 50% probability of endorsing the symptom
Discrimination parameter	How well the symptom separates participants at low and high levels of the latent trait
Harmonization	Process of data transformation permitting different sources to be treated equivalently
Item bank	Collection of the individual instrument questions or ratings along with their parameter estimates derived from IRT analyses

A two-parameter model is a common approach typically used and includes a *difficulty parameter* and a *discrimination parameter*. The difficulty parameter can be interpreted as the level of the trait at which a participant picked randomly from the study population has a 50% probability of endorsing the sign or symptom. The discrimination parameter describes how well the sign or symptom separates participants at low and high levels of the latent trait.



### ***Applications of IRT: Harmonization and Item Banks***

Throughout this work, I use the term ‘harmonization’ to refer to statistical harmonization or methods to link and equate different instruments on the same metric. The definition of harmonization is “to transform data from different sources in a way that allows them to be treated as equivalent” (45). Harmonization is not considered a technical term, while linking or test score equating are other more technical terms seen in the literature (46, 47). The harmonization analysis performed in this work will occur through the approach described in the *Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis*, created by the Agency for Healthcare Research and Quality (47). In a formal harmonization study, there is a necessity to use statistical approaches that link measurement instruments across different studies. While many potential statistical methods would be suitable, there are clear advantages to linking using latent variable techniques inherent in IRT, which will be used in my analyses (47). IRT models specify a continuous latent trait that places all items of a construct on the same metric. This facilitates comparison, and ultimately direct statistical harmonization, between different instruments, even when there are overlapping, but disjointed items across instruments or administered to patients.

Harmonization of the identified delirium identification instruments allows for the creation of an *item bank*. An item bank is a collection of the individual instrument questions or ratings along with their parameter estimates derived from IRT

analyses. Item banks have been widely used in the field of educational testing, such as the SAT and Medical College Admission Test (MCAT). These tests rely on item banks to create alternate forms of the test which are administered to varying participants on different days, and enable equivalent scoring across the alternate test forms.

More recently, IRT has been utilized within health research. There are many National Institute of Health (NIH) initiatives to develop new and well-validated measurement instruments. One example is the Patient-Reported Outcomes Measurement Information System (PROMIS) (48). PROMIS investigators used IRT methods to help build quantitative measures to evaluate patient-reported outcomes, and to generate item banks for creation of new instruments.

In this dissertation, I will utilize IRT methods to create a harmonized item bank, and to place the most commonly used and well-validated delirium identification instruments on the same metric, called the propensity to delirium. From the item bank, I will create new forms for different uses. For example, I can create short forms from the items in the item bank with high psychometric properties to accurately and efficiently screen for delirium, or long forms to confirm diagnosis or provide reference standard-type ratings for research purposes.

### ***Specific Aims***

One major problem in the identification of delirium is that there is no single agreed upon identification instrument. In fact, a 2010 study found 24 different delirium identification measures in active use (49). Since the publication of that study, multiple additional delirium identification measures have been developed. The use of numerous different measures for delirium identification presents a potential hindrance to delirium research, as well as to clinical progress in the field. Of these measures, only a few meet criteria for robustness and proper validation. This poses a major problem, since different clinicians using different instruments may not agree on whether delirium is present or not, and thus, the diagnosis of delirium may not be accurate or consistent. It is difficult to directly compare studies that use different methods for detection, since they may disagree on the prevalence and features of delirium. Thus, results found in one study may not translate directly to another study. To overcome the problem of heterogeneity in delirium identification measures, I propose the use of modern psychometric measurement techniques to create a *single harmonized item bank* of delirium case identification. This will allow for better comparison between studies, as well as the ability to combine data from multiple studies.

The overarching goal of this dissertation project is to apply advanced psychometric methods to improve the identification of delirium. This project will proceed with the following specific aims.

**Specific Aim 1.** Determine the 4 most commonly used and well-validated instruments for delirium identification through a systematic review of the medical literature, applying standardized methodologic quality ratings (*Chapter II*).

**Specific Aim 2.** Harmonize the 4 most commonly used and well-validated delirium assessment instruments to generate an item bank, which is a collection of the individual instrument questions or ratings along with their parameter estimates derived from item response theory (IRT) analyses (*Chapter III*).

**Specific Aim 3.** Explore applications of the harmonized item bank through several approaches. First, identifying different cut-points that will optimize: (a) balanced high accuracy (Youden's J-Statistic), (b) screening (sensitivity), and (c) confirmation of diagnosis (specificity) in identification of delirium. Second, comparing performance characteristics of short forms (versus long forms) developed from the item bank (*Chapter IV*).

## **CHAPTER II – Detecting Delirium: A Systematic Review of Identification Instruments for Non-ICU Settings**

Chapter II is adapted almost verbatim from a manuscript I published in the Journal of the American Geriatrics Society with permission.

### ***Abstract***

Objectives: Delirium manifests clinically in varying ways across settings. Over 40 instruments currently exist for characterizing the varying manifestations of delirium. We evaluated all delirium identification instruments according to their psychometric properties and frequency of citation in published research.

Design: We conducted the systematic review by searching CINAHL, Cochrane, EMBASE, PsycINFO, PubMed, and Web of Science from January 1, 1974- January 31, 2020, with the key words “*delirium*” and “*instruments*”, along with their known synonyms. We selected only systematic reviews, meta-analyses, or narrative literature reviews including multiple delirium identification instruments.

Measurements: Two reviewers assessed eligibility of articles and extracted data on all potential delirium identification instruments. Using the original publication on each instrument, the psychometric properties were examined using the

Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) framework.

Results: Of 2,542 articles identified, 75 met eligibility criteria, yielding 30 different delirium identification instruments. A count of citations was determined using Scopus for the original publication for each instrument. Each instrument underwent methodologic quality review of psychometric properties using COSMIN definitions. An expert panel categorized key domains for delirium identification based on criteria from the Diagnostic and Statistical Manual (DSM)-III through DSM-5. Four instruments were notable for having at least 2 of 3 of the following: citation count  $\geq 200$ , strong validation methodology in their original publication, and fulfillment of DSM-5 criteria. These were, alphabetically: Confusion Assessment Method (CAM), Delirium Observation Screening Scale (DOSS), Delirium Rating Scale-Revised-98 (DRS-R-98), and Memorial Delirium Assessment Scale (MDAS).

Conclusion: Four commonly used and well-validated instruments can be recommended for clinical and research use. An important area for future investigation is to harmonize these measures to compare and combine studies on delirium.

## ***Introduction***

Delirium is a major public health problem, impacting an estimated 2.6 million older Americans annually and accounting for over \$164 billion in healthcare expenditures (16). Delirium disproportionately affects people over age 65 and is associated with prolonged hospitalization, cognitive decline, and heightened risks for dementia and death (12, 13). Clinically, many cases of delirium go unrecognized (50), representing missed opportunities for prevention of delirium (18). A study revealed that in 61% of hospitalized patients with confirmed delirium by a palliative care expert, the diagnosis was missed by the primary referring team (19). At least in part, the lack of a unified, accepted diagnostic approach adds to the challenges of recognition (51).

The growing awareness of the seriousness of delirium, coupled with the fact that it remains a purely clinical diagnosis—without a laboratory test—has resulted in many tools for its detection. Currently, there are over 40 delirium instruments for different purposes (e.g., screening, diagnosis, and severity), targeting different clinical settings (e.g., intensive care unit (ICU), emergency department, medical wards), and intended for different users (e.g., psychiatrists, geriatricians, nurses). These instruments describe varying domains of delirium. This overabundance of instruments makes direct comparisons or interpretation of results across studies challenging.

Our overall goal was to examine instruments used for identification of delirium, defined as those used for screening or diagnosis. We aimed to conduct a comprehensive systematic review to identify the most commonly used and originally well-validated instruments for identification of delirium.

### ***Methods***

Our approach involved five steps. First, we performed a comprehensive search of the literature for reviews of delirium identification instruments from January 1, 1974 through January 31, 2020. Second, we enumerated the citations of the original publication of each instrument. Third, we evaluated the psychometric characteristics of each instrument and rated the methodologic quality of the original publication of the instrument, employing the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) framework (25, 27, 52). Fourth, we used an expert panel to identify the domains of delirium critical to identification based on Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria. Finally, the expert panel used a combination of the count of citations, the COSMIN methodologic rating, and fulfillment of DSM criteria to determine the delirium identification instruments to recommend.

Our approach to conducting and reporting of this systematic review followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines and Institute of Medicine (IOM) Standards for Systematic Reviews (53,



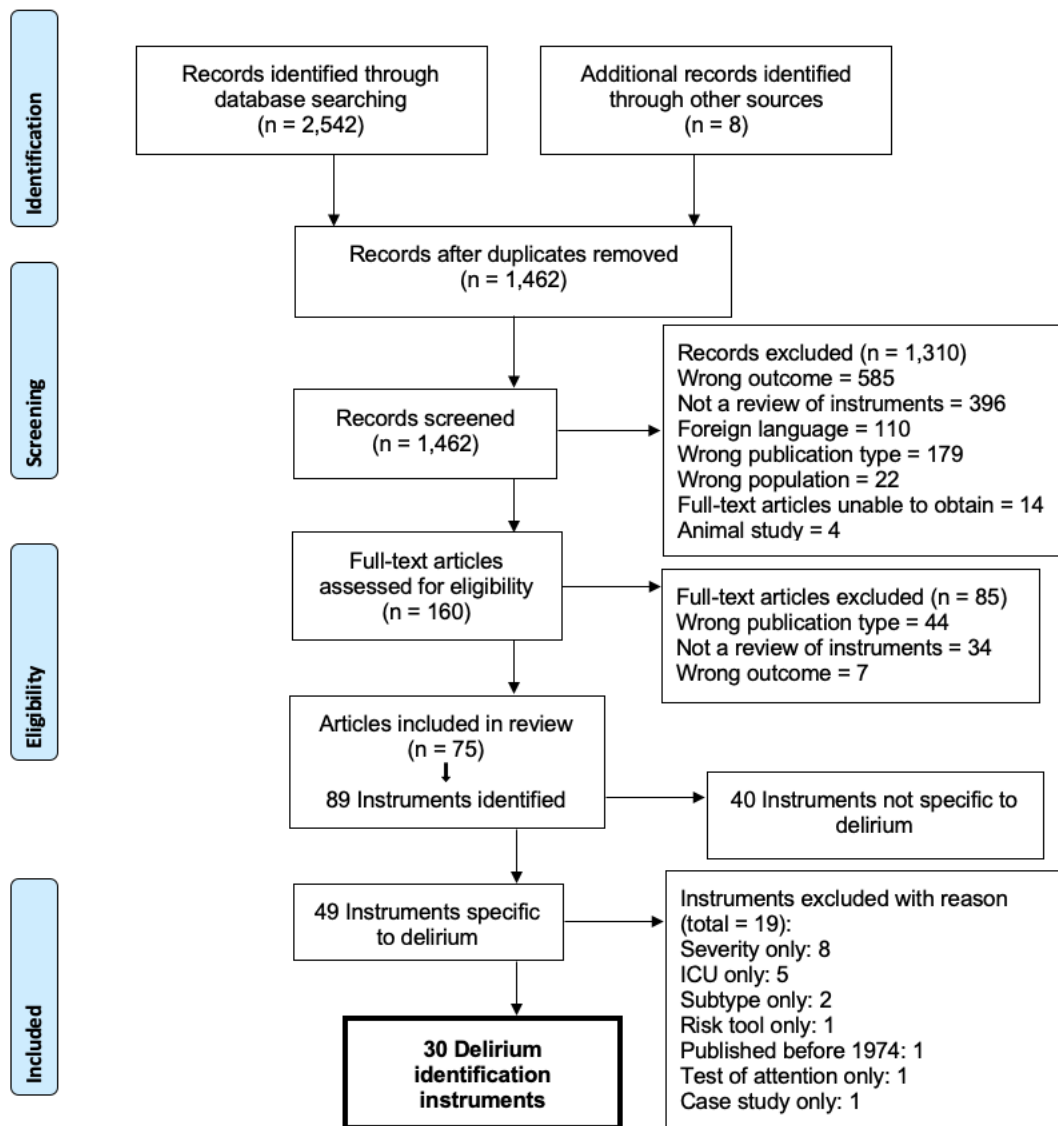
54). For the systematic review, our goal was to discover as many delirium identification instruments as possible. Since the goal of the study was to identify the most frequently cited instruments, we chose the accepted approach of a review of reviews as the most effective and efficient way to achieve this goal (55, 56). Our search began in 1974, the year that the DSM-III first codified delirium (10), and was inclusive through January 31, 2020.

### Data Sources and Searches

We identified articles through searches of 6 different databases: Cumulative Index to Nursing and Allied Health Literature (CINAHL), Cochrane Library, Excerpta Medica Database (EMBASE), PsycINFO, PubMed, and Web of Science. The search terms included the keywords “*delirium*” and “*instruments*”, along with their known synonyms (**Table 2.6**). We limited articles to review articles (systematic review, meta-analysis, or narrative review) with delirium as the main outcome. We required articles to include a minimum of two instruments. For any systematic review of a single instrument, we ensured the instrument was included in another selected article before exclusion. Exclusion criteria included studies exclusively examining alcohol-related delirium (delirium tremens), studies exclusively in pediatric populations, and other article types (i.e., case reports, commentaries, letters, editorials, conference abstracts), or studies where no full-text article was available. Because of the volume of citations to review by primary English language investigators, we restricted to English-language articles only.

Prior studies have indicated that this approach does not substantially bias systematic reviews (57). **Figure 2.1** shows the flow diagram for selection of articles. The articles underwent first-pass screening based on the title and abstract, then second-pass screening was conducted using the full-text article.

**Figure 2.1. Systematic review flow diagram**



### Title and Abstract Initial Screening

Before screening, duplicates and non-English language articles were removed by Endnote X9 software and manual cross-check. The first-pass screening of title and abstract was completed by 2 independent reviewers (B.H., M.D.) to exclude articles that did not meet eligibility criteria. Each reviewer independently reviewed the abstracts and used the RAYYAN QCRI (58) software to record results, completely blinded to the other's ratings. Articles without an abstract were included in the full-text review. If the article was rated as eligible by either of the two reviewers, the article was included for full-text review. Excluded articles were assigned a single reason for exclusion: studies restricted to pediatric populations; studies using only animal models; studies in which delirium was not the outcome; not a review; or did not evaluate at least two instruments (**Figure 2.1**).

### Full-Text Review

After the first-pass review, two independent reviewers (B.H., P.T.) established final eligibility through full-text review. If the article was rated as eligible by either of the two reviewers, the article was included for data extraction. Each rater logged their results in a Google Form in a blinded fashion. Excluded articles were given a single reason for exclusion with the same options described. Since the goal of this step was to comprehensively identify all potential delirium identification instruments, we did not conduct an appraisal of the quality of these

reviews. We used the systematic reviews, combined with hand searches of references and consultations with experts to assure comprehensive identification. Once we had found all the instruments, then the next step was to appraise the quality of the original studies of those instruments. For eligible articles, information extracted included: citation, article type (systematic review, meta-analysis, narrative review), databases and dates searched, search terms, and number of studies and instruments included in the review. Finally, to minimize biased selection based on requiring reporting in an electronic database, and as recommended by the IOM standards for systematic reviews (54), reviewers searched the reference lists of any included articles to identify other articles to include. We augmented our electronic search with hand reviews and with queries to our experts.

Our goal was to identify all potential instruments used to identify delirium. A full list of the instruments discovered from the eligible articles was presented to our expert panel. We excluded those not specific to delirium (i.e., cognitive screens, sedation instruments, dementia instruments). With the expert panel, we identified several instruments specific to delirium not found in the systematic review to bolster our final list of eligible instruments. At this stage, the experts advised excluding instruments designed solely for use in the ICU since these patients are often non-verbal, resulting in the need for unique assessments that might not be comparable with other instruments or generalizable to other settings. In addition,

a systematic review of delirium identification instruments for the ICU had been recently published (59). Since this was a study of delirium identification instruments, we chose to additionally exclude instruments measuring only severity and subtypes (hypoactive or hyperactive).

### Citation Count

We obtained the original publication for each of the eligible delirium identification instruments. The count of citations of the original publication was determined from Scopus for the date range January 1, 1974-January 31, 2020.

### COSMIN-Guided Methodologic Rating

Our goal for the second-stage review was to evaluate the psychometric characteristics of the instrument and the methodologic quality of the original publication for each selected delirium instrument. We chose the single earliest publication for each instrument. We made an exception for the Delirium Rating Scale (DRS) and used the later study since the instrument had been revised [Delirium Rating Scale-Revised-98 (DRS-R-98)]. We rated the Confusion Assessment Method (CAM) long form and short form separately. A single publication per instrument was used to minimize bias as older instruments might have multiple validation studies. Our quality rating was based on an approach we published previously (**Table 2.7**) (9). Our approach used the COSMIN standards of measurement properties (25, 27, 52). The COSMIN rating was utilized to

evaluate the psychometric properties of the instrument as reported in its original study. Each article was reviewed independently in a blinded fashion by at least two of three reviewers (B.H., K.E., J.Y.) and rated according to the COSMIN framework. The assessment items include ratings of published descriptions of effect indicators, internal consistency, content validity, inter-rater reliability, construct/convergent validity, and criterion validity (full definitions and scoring are in **Table 2.7**). Estimates and sample sizes for these different types of reliability and validity were recorded. The few small differences between the two independent COSMIN ratings of each article were adjudicated by a third rater (R.N.J.).

The ratings on each of the COSMIN criteria were summed and reported as a 0 to 6 score (**Table 2.7**), using an adaptation of the COSMIN scoring procedure published previously (9, 26). For reporting on each of these categories the instruments were given one point; failure to report on these categories resulted in no points. If a category was reported, but used sample sizes less than 50, only a half point was assigned.

#### Expert Panel Review of Instruments

We assembled an interdisciplinary expert panel to determine the key domains for identification of delirium and ascertained their alignment with DSM criteria. Experts from geriatric medicine (S.K.I., T.T.H., one anonymous), geriatric

psychiatry (E.D.M.), cognitive neurology (T.G.F.), gerontological nursing (P.T.), and social work (E.M.S.) were included in the panel. Face-to-face meetings were done twice in consensus sessions following a modified Delphi approach (35, 60) to adjudicate the criteria, with independent, blinded ranking assignments between meetings. We reviewed criteria enumerated in DSM-III, DSM-III-R, DSM-IV, DSM-IV-TR, and DSM-5 (10, 11, 61-63). Each individual criterion was first assigned to domain(s) identified previously (9, 64). Then, the expert panel rated whether each domain was essential for delirium identification; consensus was considered achieved with agreement by 6/7 (86%). The expert panel determined whether each of the 30 delirium identification instruments fulfilled DSM-5 criteria.

Subsequently, the expert panel determined the criteria for selecting the instruments to recommend. After consensus, the following criteria were selected: citation count  $\geq 200$ , COSMIN score  $> 4$ , and meeting full DSM-5 criteria. To be recommended, an instrument should meet at least 2 of these 3 criteria.

## ***Results***

Results of the systematic review are shown in **Figure 2.1**. The literature review yielded 2,542 articles, which were narrowed based on our exclusion criteria to 160 articles for full-text review. From full-text review, 75 articles (47%) met our inclusion criteria (**Table 2.8**). We identified 89 total instruments. The expert panel

determined 49 were specific to delirium; we excluded 19 for the following reasons: measuring severity only (n=8); intended for ICU patients (n=5); measuring only delirium subtypes (hypoactive or hyperactive) (n=2); measuring only risk for developing delirium (n=1); including only attention tests (n=1); published before 1974 (n=1); and case report only (n=1) (**Table 2.9**). Thus, our study included 30 delirium-specific identification instruments developed for use in non-ICU settings (**Table 2.10**). Of these 30 instruments, allowing for multiple categories, usage was 87% for screening, 27% for diagnosis, and 10% for severity. The most common study populations examined included: medical and/or surgical wards (47%), geriatric wards (20%), emergency department (10%), and long-term care facilities (10%). The reference standard used for each study included: DSM (40%), CAM (20%), expert clinical judgment only (13%), and not described or not used (27%).

**Table 2.1** shows characteristics of the full-text articles reviewed. There were 18 articles that mentioned at least 10 instruments. No articles were published before 1990, however, since that time article count has risen exponentially. The 75 included articles individually reviewed between 2 and 19,000 articles.



**Table 2.1. Characteristics of articles reviewed**

<b>Characteristic</b>	<b>N</b>	<b>%</b>
<b>Number of instruments described</b>	<b>75</b>	<b>100</b>
<b>(n, %)</b>		
2	9	12
3	4	5
4	6	8
5	13	18
6	8	11
7	7	9
8	7	9
9	3	4
10-14	8	11
15-19	4	5
≥20	6	8
<b>Year published (n, %)</b>		
1974-1989	0	0
1990-2000	5	7
2001-2010	17	23
2011-2014	25	33
2015-2019	28	37

**Article Type (n, %)**

Meta-analysis	5	7
Systematic review	23	30
Narrative review	47	63

---

**Table 2.2** shows the selection criteria for all the delirium identification instruments. Four instruments stand out for satisfying most of the COSMIN framework criteria, assessing many of the DSM-5 criteria, and widespread use as evidenced by their high citation count. These were the Confusion Assessment Method (CAM) [2,685 citations, COSMIN criteria count = 4.5, full DSM-5 criteria], Delirium Rating Scale-Revised-98 (DRS-R-98) [499 citations, COSMIN criteria count = 4.5, full DSM-5 criteria], Memorial Delirium Assessment Scale (MDAS) [492 citations, COSMIN criteria count = 5, partial DSM-5 criteria], and the Delirium Observation Screening Scale (DOSS) [212 citations, COSMIN criteria count = 6, partial DSM-5 criteria].

**Table 2.2. Selection criteria for delirium identification instruments based on the original citation**

<b>Name of Scale</b>	<b>*Count of Citations (Scopus: January 1, 1974- January 31, 2020)</b>	<b>COSMI N Score (Max=6)</b>	<b>Meets DSM-5 Criteria (Yes/No)</b>
Confusion Assessment Method (CAM) - Long Form and Short Form	2909	4.5	Yes
Delirium Rating Scale-Revised-98 (DRS-R-98)	552	4.5	Yes
Memorial Delirium Assessment Scale (MDAS)	532	5	No
Delirium Observation Screening Scale (DOSS)	238	6	No
Chart Delirium Identification (CHART-DEL)	216	3.5	No
Neelon and Champagne confusion scale (NEECHAM)	207	5	No
Delirium Symptom Interview (DSI)	204	4	No
Confusion Assessment Method Emergency Department (CAM-ED)	176	2.5	No
4 "A"s test (4AT)	168	4	No
Delirium Triage Screen (DTS)	117	4	No
Brief Confusion Assessment Method (bCAM)	117	4	No
3-Minute Diagnostic Assessment (3D-CAM)	98	4	Yes
Saskatoon Delirium Checklist (SDC)	97	2	No
Single Question in Delirium (SQiD)	64	2	No
Nursing Home-Confusion Assessment Method (NH-CAM)	58	2	Yes
Family-Confusion Assessment Method (FAM-CAM)	48	3.5	Yes

Clinical Assessment of Confusion-A (CAC-A)	40	3.5	No
Recoverable Cognitive Dysfunction Scale (RCDS)	34	2	No
modified Confusion Assessment Method for the Emergency Department (mCAM-ED)	28	3	No
Delirium Diagnostic Tool-provisional (DDT-Pro)	27	3.5	No
Confusion Rating Scale (CRS)	26	4	No
Bedside Confusion Scale (BCS)	25	2.5	No
Nursing Delirium Screening Scale (Nu-DESC)	24	2.5	No
Recognizing Acute Delirium as Part of Your Routine (RADAR)	24	4	No
Visual Analog Scale for Acute Confusion (VAS-AC)	22	3	No
Inter Resident Assessment Instrument Acute Care (InterRAI AC)	13	4	No
Simple Query for Easy Evaluation of Consciousness (SQeeC)	10	4	No
Informant Assessment of Geriatric Delirium scale (I-AGeD)	9	4.5	No
Clinical Assessment of Confusion-B (CAC-B)	NA	3.5	No
Organic Brain Syndrome (OBS)	NA	NR	NR

---

\*Descending order by count of citations.

Abbreviations: COSMIN, Consensus-based Standards for the Selection of Health Measurement Instruments; DSM, Diagnostic and Statistical Manual; NA, Not attainable; NR, No Rating

---

**Figure 2.2** shows the domain coverage of the CAM, DOSS, DRS-R-98, and MDAS. Domains covered by each instrument were classified as fulfilling DSM-5 criteria, other DSM diagnostic criteria, or other associated features. They are listed in descending order by number of total domains covered, with the DRS-R-98 assessing 13 domains, the CAM long form assessing 11 domains, the MDAS assessing 10 domains, and the DOSS assessing 9 domains. The CAM short form overlaps with the CAM long form and was excluded from this analysis. For the DSM-5 criteria, all instruments included core criteria of inattention, disorientation, and cognitive impairment; however, two instruments (MDAS and DOSS) did not include acute onset and fluctuating course. In other DSM criteria, all 4 overlapped with the same domains on 4/6 criteria (disorganized thinking, psychomotor agitation, psychomotor retardation, and hallucinations), and all but the DRS-R-98 included altered level of consciousness. Only the DRS-R-98 included organic etiology.

**Figure 2.2. Domain coverage of 4 recommended delirium instruments**

Instrument	DSM-5 criteria			Other DSM diagnostic criteria				Other associated features			Number of Domains Assessed				
	Acute onset	Fluctuating course	Inattention	Disorientation	Cognitive impairment	Disorganized thinking	Level of consciousness	Psychomotor agitation	Psychomotor retardation	Hallucinations, perceptual disorder or distortion		Organic Etiology	Delusions	Sleep disturbance	Emotional lability
Delirium Rating Scale Revised-98 (DRS-R-98)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	13
Confusion Assessment Method (CAM) - Long Form	●	●	●	●	●	●	●	●	●	●		●	●	●	11
Memorial Delirium Assessment Scale (MDAS)			●	●	●	●	●	●	●	●		●	●	●	10
Delirium Observation Scale (DOSS)			●	●	●	●	●	●	●	●				●	9

**Table 2.3** shows a comparison of the CAM, DOSS, DRS-R-98 and MDAS. These instruments had the highest citation count and COSMIN score. We also show the number of DSM-5 criteria and delirium identification domains met by each of the top 4 instruments. **Table 2.3** provides additional information about these instruments including time for completion, qualifications of the raters, and evidence of construct and criterion validity. Notably, each of the instruments used a reference standard delirium diagnosis by a physician based on DSM criteria. Full details of the review of COSMIN criteria and other details for each instrument is described in **Table 2.4** and **Table 2.5**.

**Table 2.3. Comparison of 4 recommended delirium instruments (alphabetical order)**

<b>Delirium Instrument, year of publication, (Sample size)</b>	<b>Recommended Time to Complete</b>	<b>Qualifications of Raters (original study)</b>	<b>Construct Validity<sup>a</sup></b>	<b>Criterion Validity<sup>b</sup></b>	<b>COSMIN Rating, (best=6)</b>	<b>Citations (Scopus)</b>	<b>Number of DSM-5 criteria fulfilled</b>	<b>Domains Covered, Number</b>
Confusion Assessment Method (CAM), 1990 (N = 56)	10-15 minutes (long form), 3-5 minutes (short form)	Trained lay or clinical raters	r=.64 with MMSE r=.59 with story recall r=.82 with VAS-C r=.66 digit span	DSM-III-R criteria by psychiatrist	4.5	2909	5/5	11
Delirium Observation Scale (DOSS), 2003 (N = 92)	< 5 minutes	Nurses without specialized training	r=.60-.79 with MMSE r=.63 with CAM r=.33-.74 with IQCODE	DSM-IV criteria by geriatrician	6	238	3/5	9
Delirium Rating Scale-Revised-98 (DRS-R-98), 2001 (N= 26)	20-30 minutes (scoring), following ~1 hour (gathering information from nurse, family, chart)	Psychiatrically trained clinicians	r=.41 with CTD	DSM-IV criteria by referring service physician	4.5	552	5/5	13
Memorial Delirium Assessment Scale (MDAS), 1997 (N = 30)	10-15 minutes (scoring), following 15-30 minutes (interview, information from nurse, family, chart)	Trained clinicians	r=.91 with MMSE r=.89 with CGR r=.88 with DRS	DSM-III-R or DSM-IV criteria by psychiatrist	5	532	3/5	10

---

Abbreviations: CGR, Clinician's Global Rating; CTD, Cognitive Test for Delirium; DRS, Delirium Rating Scale; DSM, Diagnostic and Statistical Manual; COSMIN, Consensus-based Standards for the Selection of Health Measurement Instruments; MMSE, Mini-Mental State Examination; VAS-C, Visual Analog Scale for Confusion.

<sup>a</sup>Construct validity represents a test of correlations with other instruments of the same construct, in this case delirium identification. For  $r$ ,  $>0.7$  indicates a strong relationship,  $>0.5$  indicates a moderate relationship, and  $>0.3$  indicates a weak relationship.

<sup>b</sup>Criterion validity represents the reference standard assessment used.

---



**Table 2.4. COSMIN score of delirium identification instruments**

<b>Instrument &amp; citation</b>	<b>Quality Rating (max: 6)</b>	<b>Points, up to 1 for each category (content, effect indicators, internal consistency, inter-rater, construct, external validity; see text for details)</b>
<b>Delirium Observation Screening Scale (DOSS)</b>	<b>6</b>	
<b>Memorial Delirium Assessment Scale (MDAS)</b>	<b>5</b>	-1/2 for fair sample size or smaller (<50) for internal consistency reliability -1/2 for fair sample size or smaller (<50) for inter-rater reliability
<b>Neelon and Champagne confusion scale (NEECHAM)</b>	<b>5</b>	-1 not all effect indicators
<b>Confusion Assessment Method (CAM) – Long Form and Short Form</b>	<b>4.5</b>	-1 No internal consistency reliability -1/2 for fair sample size or smaller (<50) for inter-rater reliability
<b>Delirium Rating Scale Revised-98 (DRS-R-98)</b>	<b>4.5</b>	-1/2 for fair sample size or smaller for internal consistency reliability -1/2 for fair sample size or smaller (<50) for inter-rater reliability -1/2 for fair sample size or smaller (<50) for criterion validity
<b>Informant Assessment of Geriatric Delirium scale (I-AGeD)</b>	<b>4.5</b>	-1 No interrater reliability -1/2 for fair sample size or smaller (<50) for construct validity
<b>Confusion Rating Scale (CRS)</b>	<b>4</b>	-1 No internal consistency reliability -1 No external validation
<b>3-Minute Diagnostic Assessment (3D-CAM)</b>	<b>4</b>	-1 No internal consistency -1 No construct validity
<b>4 "A's" Test (4AT)</b>	<b>4</b>	-1 No interrater reliability -1 No construct validity
<b>Brief Confusion Assessment Method (b-CAM)</b>	<b>4</b>	-1 No internal consistency -1 No construct validity

<b>Delirium Triage Screen (DTS)</b>	4	-1 No internal consistency -1 No construct validity
<b>Delirium Symptom Interview (DSI)</b>	4	-1 not all effect indicators -1 No construct validity
<b>Inter Resident Assessment Instrument Acute Care (InterRAI AC)</b>	4	-1 No internal consistency -1 No construct validity
<b>Recognizing Acute Delirium as Part of Your Routine (RADAR)</b>	4	-1 No content validity -1 No internal consistency
<b>Simple Query for Easy Evaluation of Consciousness (SQeeC)</b>	4	-1 No internal consistency -1 No interrater reliability
<b>Clinical Assessment of Confusion-A (CAC-A)</b>	3.5	-1 No internal consistency -1/2 for fair sample size or smaller (<50) for inter-rater reliability -1 No criterion validity
<b>Clinical Assessment of Confusion-B (CAC-B)</b>	3.5	-1/2 small sample size (<50) for inter-rater reliability -1 No construct validity -1 No criterion validity
<b>Chart Delirium Identification (CHART-DEL)</b>	3.5	-1 No internal consistency -1/2 interrater reliability sample size and methods not reported -1 No construct validity
<b>Delirium Diagnostic Tool-provisional (DDT-Pro)</b>	3.5	-1 No internal consistency -1/2 inter-rater reliability sample size not reported -1/2 for fair sample size or smaller (<50) for construct validity -1/2 for fair sample size or smaller (<50) for criterion validity
<b>Family-Confusion Assessment Method (FAM-CAM)</b>	3.5	-1 No internal consistency -1/2 inter-rater reliability sample size not reported -1 No construct validity

<b>Modified confusion assessment method for the ED (mCAM-ED)</b>	<b>3</b>	-1 No internal consistency -1 No interrater reliability -1 No construct validity
<b>Visual Analog Scale for Acute Confusion (VAS-AC)</b>	<b>3</b>	-1 uncertain effect indicators -1 No internal consistency -1/2 for fair sample size or smaller (<50) for interrater reliability -1/2 No correlations given for construct validity
<b>Nursing Delirium Screening Scale (Nu-DESC)</b>	<b>2.5</b>	-1 No internal consistency -1 No inter-rater reliability -1/2 unclear sample size reporting (construct validity) -1 No external validation
<b>Bedside Confusion Scale (BCS)</b>	<b>2.5</b>	-1 No internal consistency -1 No inter-rater reliability -1/2 for fair sample size or smaller (<50) -1 No construct validity
<b>Confusion Assessment Method-Emergency Department (CAM-ED)</b>	<b>2.5</b>	-1 No internal consistency -1 No interrater reliability -1 No construct validity -1/2 for fair sample size or smaller (<50) for criterion validity
<b>Recoverable Cognitive Dysfunction Scale (RCDS)</b>	<b>2</b>	-1 Content validity not discussed -1 No internal consistency -1 No interrater reliability -1 No external validity
<b>Nursing Home Confusion Assessment Method (NH-CAM)</b>	<b>2</b>	-1 uncertain effect indicators -1 No internal consistency -1 No construct validity -1 No criterion validity

**Saskatoon Delirium Checklist (SDC)**

**2**

- 1 No internal consistency
- 1 No inter-rater reliability
- 1 No construct validity
- 1 No criterion validity

---

**Single Question in Delirium (SQiD)**

**2**

- 1 uncertain content validity
- 1 No internal consistency reported
- 1 No interrater reliability
- 1/2 for fair sample size or smaller (<50) for construct validity
- 1/2 for fair sample size or smaller (<50) for criterion validity

---

**Organic Brain Scale (OBS)**

**NR**

---

Organic Brain Scale original article could not be obtained to rate; NR = no rating

**Table 2.5. Comparison of delirium instruments (alphabetical order)**

<b>Delirium Instrument</b>	<b>Recommended Time to Complete</b>	<b>Qualifications of Raters (original study)</b>	<b>Construct Validity<sup>a</sup></b>	<b>Criterion Validity<sup>b</sup></b>	<b>Number of items</b>	<b>COSMIN Rating, (best=6)</b>	<b>Citations (Scopus)</b>
Bedside Confusion Scale (BCS)	2 minutes to administer and rate	Not reported	NR	CAM	2	2.5	25
Brief Confusion Assessment Method (bCAM)	~1 minute	Trained lay raters	NR	DSM-IV criteria by psychiatrist	4	4	117
Chart Delirium Identification (CHART-DEL)	Not reported	Trained nurses	NR	Interviewer rating using the CAM	7	3.5	216
Clinical Assessment of Confusion-A (CAC-A)	Not reported	Not reported	53% VAS-C	NR	25	3.5	40
Clinical Assessment of Confusion-B (CAC-B)	Not reported	Nurses	NR	NR	58 items with 7 sub scales	3.5	NA
Confusion Assessment Method (CAM)	10-15 minutes (long form), 3-5 minutes (short form)	Trained lay or clinical raters	r=.64 with MMSE r=.59 with story recall r=.82 with VAS-C r=.66 digit span	DSM-III-R criteria by psychiatrist	Long form = 9 Short form = 4	4.5	2909
Confusion Assessment Method	30 minutes, including MMSE, 5 minutes to	Trained nurses working in the	NR	CAM	MMSE + 10 items	2.5	176

Emergency Department (CAM-ED)	complete CAM portion	emergency department						
Confusion Rating Scale (CRS)	Not reported	Nurses	r = .51 SPMSQ	NR	4	4	26	
Delirium Diagnostic Tool-provisional (DDT-Pro)	Not reported	Not reported	r = 0.889 DRS-R-98	DSM-IV-TR criteria rated by a clinical neuropsychologist	3	3.5	27	
Delirium Observation Scale (DOSS)	< 5 minutes	Nurses without specialized training	r=.60-.79 with MMSE r=.63 with CAM r=.33-.74 with IQCODE	DSM-IV criteria by geriatrician	Original 25-item form, revised 13-item form	6	238	
Delirium Rating Scale-Revised-98 (DRS-R-98)	20-30 minutes (scoring), following ~1 hour (gathering information from nurse, family, chart)	Psychiatrically trained clinicians	r=.41 with CTD	DSM-IV criteria by referring service physician	Total = 16 Severity only = 13	4.5	552	
Delirium Symptom Interview (DSI)	10-15 minutes for interview	Clinicians or lay raters	NR	Clinical judgment of psychiatrist and neurologist	32	4	204	
Delirium Triage Screen (DTS)	~20 seconds	Trained lay raters	NR	DSM-IV criteria by psychiatrist	2	4	117	
Family-Confusion Assessment Method (FAM-CAM)	5-10 minutes	Trained lay or clinical raters	NR	CAM rated by trained research assistants	11	3.5	48	

Informant Assessment of Geriatric Delirium scale (I-AGeD)	Not reported	Trained caregivers	r = .28-48 DOS Sens: 81.5%, Spec: 64.4% CAM	DSM-IV criteria by geriatric residents	10	4.5	9
Inter Resident Assessment Instrument Acute Care (InterRAI AC)	Not reported	Trained nurses	NR	DSM-IV criteria by geriatrician	4	4	13
Memorial Delirium Assessment Scale (MDAS)	10-15 minutes (scoring), following 15-30 minutes (interview, information from nurse, family, chart)	Trained clinicians	r=.91 with MMSE r=.89 with CGR r=.88 with DRS	DSM-III-R or DSM-IV criteria by psychiatrist	10	5	532
modified Confusion Assessment Method for the Emergency Department (mCAM-ED)	One minute to rate attention and 3 to 5 minutes to complete the assessment	Nurses and trained lay interviewers	NR	Senior emergency physician	Not reported	3	28
Neelon and Champagne confusion scale (NEECHAM)	10 minutes to rate, including measuring vital signs	Nurses	r = .87 MMSE	DSM-III-R criteria by trained research nurse	9	5	207
Nursing Delirium Screening Scale (Nu-DESC)	~1 minute	Nurses	r = .71 DSM-IV r = .67 MDAS	NR	5	2.5	24
Nursing Home-Confusion Assessment	Not reported	Nursing home staff	NR	NR	9	2	58

Method (NH-CAM)							
Organic Brain Syndrome (OBS)	NA	NA	NR	NR	NA	NR	NA
Recognizing Acute Delirium as Part of Your Routine (RADAR)	<1 minute to score, average of 7 seconds	Nursing or other clinical staff, can be rated by trained lay rater	52% to 85% agreement with CAM symptoms	DSM-IV-TR diagnostic criterion completed by trained research assistant	3	4	24
Recoverable Cognitive Dysfunction Scale (RCDS)	Not reported	Not reported	Kappa = 0.93 with CAM, DRS, CAMDEX	DSM-III-R; ICD-10; CAMDEX	4	2	34
Saskatoon Delirium Checklist (SDC)	15 minutes to administer	Not reported	NR	NR	9	2	97
Simple Query for Easy Evaluation of Consciousness (SQeeC)	30 seconds to 1 minute	Not reported	Sensitivity of 83%, specificity of 81%; SqID sensitivity 77%, specificity 51%	DSM-IV criteria by geriatrician	2 items, 4 questions	4	10
Single Question in Delirium (SQiD)	Not reported	Not reported	Performed, but no correlations reported	Psychiatrists interview	1	2	64
Visual Analog Scale for Acute Confusion (VAS-AC)	Not reported	Master's level nurses	Performed, but no correlations reported	DSM-IV criteria by investigator	Not reported	3	22



3-Minute Diagnostic Assessment (3D-CAM)	3 minutes to rate	Trained lay raters or clinicians	NR	DSM-IV criteria by clinical psychologists and practice nurses	10 interview questions, 10 observational items, 2 supplementary questions	4	98
4 "A"s test (4AT)	<2 minutes including brief cognitive testing embedded in the interview	Clinicians	NR	Geriatrician diagnosis using DSM-IV-TR criteria	4	4	168

Abbreviations: CAMDEX, Cambridge Mental Disorders of the Elderly Examination; CGR, Clinician's Global Rating; CTD, Cognitive Test for Delirium; DRS, Delirium Rating Scale; DSM, Diagnostic and Statistical Manual; COSMIN, Consensus-based Standards for the Selection of Health Measurement Instruments; MMSE, Mini-Mental State Examination; NA, not attainable; NR, no rating; SPMSQ, the short portable mental status questionnaire; VAS-C, Visual Analog Scale for Confusion.

<sup>a</sup>Construct validity represents a test of correlations with other instruments of the same construct, in this case delirium identification. For  $r$ ,  $>0.7$  indicates a strong relationship,  $>0.5$  indicates a moderate relationship, and  $>0.3$  indicates a weak relationship.

<sup>b</sup>Criterion validity represents the reference standard assessment used.

## ***Discussion***

The ability to accurately identify delirium is important to provide optimal clinical care. Moreover, to advance the field, it is critical to have reliable approaches for delirium identification. We identified 30 delirium identification instruments used in non-ICU settings. We evaluated several aspects of each instrument including citation count, satisfaction of COSMIN criteria for the evaluation of health measurement instruments, and expert panel guidance regarding the coverage of DSM-5 criteria for delirium. Based on our systematic review combined with an expert panel process, we recommend (in alphabetical order) the CAM, DOSS, DRS-R-98, and MDAS as frequently used and well-validated instruments to identify delirium that are at least partially consistent with the current diagnostic framework (DSM-5) for delirium.

Each of these instruments identifies delirium somewhat differently, assessing different domains. Each was designed for use by different users in varying clinical settings. Thus, the choice in selecting an instrument to identify delirium should be guided by these factors along with logistical considerations for the intended clinical or research application. While different instruments may be preferred for clinical versus research uses, both settings seek approaches to maximize reliability, validity, and minimize costs and burden of assessment. However, in the clinical setting, users often prioritize expediency, which may be counter-balanced by sub-optimal diagnostic accuracy.

For the selected instruments, to assist nurses in rapid delirium identification during each shift, the DOSS provides a brief (<5 minute) rating with minimal training. Although the ratings gather important information assessing clinical progress, an experienced clinician is required to confirm and establish diagnoses. Use of the DRS-R-98 may be preferred by skilled psychiatrically-trained clinicians since it provides detailed ratings and has been used in phenomenological delirium studies. However, the administration of the DRS-R-98 is time consuming (20-30 minutes) and labor intensive compared. The MDAS is scored with or without additional tests such as the Mini-Mental State Examination (24). However, all three of these instruments have no built-in diagnostic algorithm, and use cut-points to identify delirium. Thus, a delirium diagnosis can be achieved with multiple different domains.

The CAM can be rated by trained lay interviewers, nurses, or physicians. Scored according to a diagnostic algorithm, the CAM aligns with the DSM-5 diagnostic criteria. There are two forms, a short-form which allows rapid assessment (<5 minute) and a long-form (10-15 minutes) to help establish diagnoses in clinical and research applications. The availability of two different forms may offer advantages for large-scale clinical applications or studies. The CAM has been integrated into numerous electronic medical record systems. While the CAM

short-form is widely used as a reliable screening instrument (1, 65, 66), it does not cover as many domains as the other selected instruments.

Our work extends the findings of two previous reviews. Adamis and colleagues used extensive search strategies to define the features of 24 different delirium instruments, including their psychometric properties (49), which were rated on a scale from +++ to -. This review did not utilize a uniform approach to characterize psychometric properties reported across studies. They recommended the CAM, DRS, MDAS, and Neelon and Champagne Confusion Scale (NEECHAM) due to their robustness and ease of use. Our work extends this article by updating the search and instruments included over the past decade, and providing a more systematic approach to scoring psychometric and methodologic properties. Subsequently, van Velthuisen and colleagues used an extensive search strategy to find 28 different delirium instruments (67). Any study that described psychometric properties of delirium identification instruments was included. The studies were restricted to those that included reference standard delirium diagnoses made by a physician using the DSM, editions III, IV or 5 or the International Classification of Diseases (ICD-10). Their quality assessment was guided by QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies) (68), which assesses 4 domains including patient selection, index test, reference standard, and flow and timing. The psychometric properties included in their review included sensitivity and specificity, inter-rater reliability, and internal

consistency reliability. They recommended the CAM and Nursing Delirium Screening Scale (Nu-DESC), and the DOSS, DRS-R-98, and CAM-Intensive Care Unit (CAM-ICU) were mentioned. Our study extends this previous work by considering citation counts, aligning the instruments with DSM criteria and addressing other aspects of validity.

There are several strengths to the present study. We used rigorous approaches, including PRISMA and IOM guidelines, to guide our comprehensive systematic review. We included a count of citations of the original publication of each instrument, along with methodologic quality ratings based on the COSMIN approach. We used an expert panel process to determine the domains for delirium identification, and applied them to each instrument item. A major strength includes our review of every DSM delirium criterion since the original codification of delirium in DSM-III. By reviewing each version, we were able to identify an inclusive consensus listing of domains pertinent to delirium identification. This allowed for each version of DSM to be included, many of which served as the reference standards in the original publication. We further aligned each of our recommended instruments with the diagnostic criteria of the current DSM-5. We followed IOM guidelines to ensure instruments were not missed by including hand searches and consulting with experts about other potential instruments to include (54).

Several limitations deserve comment. First, there is a potential bias as one of the authors (S.K.I.) is a creator of four delirium identification instruments found in our review [CAM, Chart Delirium Identification (CHART-DEL), Family-Confusion Assessment Method (FAM-CAM), and 3-Minute Diagnostic Assessment (3D-CAM)]. Additionally, coauthors E.D.M and R.N.J. are creators of the 3D-CAM. We minimized bias by not including any of these coauthors in the direct COSMIN review of any instruments. Second, restricting the COSMIN review to the original publication of each instrument poses another potential limitation. It is possible that had we probed the literature for validation studies for each instrument, we could have amassed more evidence for each instrument. Third, we understand that using citation count could potentially bias towards older instruments, however, this was only one of three criteria that the expert panel selected to rank the quality of the instruments, the other two—COSMIN score and DSM-5 criteria—would not be biased by the age of the instrument. Fourth, we only considered the presence or absence of a validity or reliability assessment in an original instrument publication as a marker of the rigor of the original presentation. Our ranking may have been more precise if we had incorporated actual values of statistics used in the evaluation. However, not all studies reported all or the same statistics, used samples representative of different populations, and used different reference standards. These differences led us to take a very coarse approach to ranking the rigor of the original publication. Fifth, for reasons described earlier, we did not include instruments that were developed

for ICU patients. We acknowledge that this systematic review is not generalizable to the ICU setting. Finally, the ability to distinguish delirium in persons with underlying dementia is an area of paramount importance for future investigation. Future work will be needed to rate and rank delirium identification instruments for their ability to differentiate delirium and dementia or to identify delirium superimposed on dementia.

This study provides a broad overview of delirium identification instruments. We found numerous instruments used in different clinical settings by different raters. We were unable to recommend a single instrument for universal use, however, we found 4 instruments that are widely used and were well-validated in their original publications with a wide-range of clinical and research applications. The study helped to refine the construct of delirium through alignment of the delirium assessment items, DSM diagnostic criteria, and other previously identified delirium domains. While many studies have been published using different delirium identification instruments, comparing these studies is difficult due to the measurement heterogeneity. An important area for future investigation will be to harmonize these measures, which may help to compare results across studies, and to combine results from existing studies to form large datasets exploring pathophysiology and treatment. We hope this work will help to unify the field around delirium identification, and lay a foundation to advance the field.

**Table 2.6. Search strategies for databases****PubMed****Platform:** PubMed, 1946-Present**Year Limits:** 1974-Present**Other Limits:** Language filter: English

#1	"Delirium"[MeSH] OR "Delirium"[tiab] OR "Acute confusion"[tiab] OR "Acute organic brain syndrome"[tiab] OR "Acute confusional state"[tiab] OR "Acute brain syndrome"[tiab] OR "Acute brain failure"[tiab] OR "Acute brain dysfunction"[tiab] OR "Acute organic psychosyndrome"[tiab] OR "Acute organic psycho-syndrome"[tiab] OR "Acute psycho-organic syndrome"[tiab] OR "Acute psychoorganic syndrome"[tiab] OR "Metabolic encephalopathy"[tiab] OR "Clouded state"[tiab] OR "Clouding of consciousness"[tiab]
#2	"Psychiatric Status Rating Scales"[MeSH] OR "Neuropsychological Tests"[MeSH] OR "Psychometrics"[MeSH] OR "Mass Screening"[MeSH] OR "Geriatric Assessment"[MeSH] OR "Psychological Tests"[MeSH] OR "Surveys and Questionnaires"[MeSH] OR "Interview, Psychological"[MeSH] OR "Mental Status Schedule"[MeSH] OR "Qualitative Research"[MeSH] OR "Checklist"[MeSH] OR "Scale"[tiab] OR "Scales"[tiab] OR "Instrument"[tiab] OR "Instruments"[tiab] OR "Measure"[tiab] OR "Measures"[tiab] OR "Questionnaire"[tiab] OR "Questionnaires"[tiab] OR "Interview"[tiab] OR "Interviews"[tiab] OR "Evaluation"[tiab] OR "Evaluations"[tiab] OR "Examination"[tiab] OR "Examinations"[tiab] OR "Exam"[tiab] OR "Exams"[tiab] OR "Test"[tiab] OR "Tests"[tiab] OR "Screening"[tiab] OR "Screenings"[tiab] OR "Assessment"[tiab] OR "Assessments"[tiab] OR "Index"[tiab] OR "Indices"[tiab] OR "Indexes"[tiab] OR "Qualitative Research"[tiab] OR "Qualitative Study"[tiab] OR "Qualitative Studies"[tiab] OR "Checklist"[tiab] OR "Checklists"[tiab]
#3	"Adult"[MeSH] OR "Young Adult"[MeSH] OR "Aged"[MeSH] OR "Aged, 80 and over"[MeSH] OR "Frail Elderly"[MeSH] OR "Adult"[tiab] OR "Adults"[tiab] OR "Young Adult"[tiab] OR "Young Adults"[tiab] OR "Middle age"[tiab] OR "Middle aged"[tiab] OR "Elderly"[tiab] OR "Elder"[tiab] OR "Oldest old"[tiab] OR "Nonagenarian"[tiab] OR "Nonagenarians"[tiab] OR "Octogenarian"[tiab] OR "Octogenarians"[tiab] OR "Centenarian"[tiab] OR "Centenarians"[tiab] OR "Frail"[tiab]
#4	"Alcohol withdrawal delirium"[MeSH] OR "Alcohol withdrawal syndrome"[tiab] OR "Delirium tremens"[tiab] OR "Alcohol withdrawal delirium"[tiab]



#5	"review"[Publication Type] OR "review literature as topic"[MeSH Terms] OR "systematic review"[All Fields] OR "meta-analysis"[Publication Type] OR "meta-analysis as topic"[MeSH Terms] OR "meta-analysis"[All Fields]
#6	#1 AND #3
#7	#6 NOT #4
#8	#7 AND #5

**Embase****Platform:** Elsevier, 1947-Present; Ovid, 1988-present**Year Limits:** [01/01/1974]/sd; present**Other Limits:** English language

1	delirium/ or intensive care psychosis/ or postoperative delirium/
2	(delirium or acute confusion or acute organic brain syndrome or acute confusional state or acute brain syndrome or acute brain failure or acute organic psychosyndrome or acute psycho-organic syndrome or metabolic encephalopathy or clouded state or clouding of consciousness).tw.
3	or/1-2
4	psychological rating scale/ or psychometry/ or mass screening/ or geriatric assessment/ or exp psychologic test/ or exp questionnaire/ or qualitative research/ or rating scale/ or clinical assessment tool/ or clinical assessment/ or exp interview/ or clinical evaluation/ or screening test/
5	(mental status schedule or scale or scales or instrument or instruments or measure or measures or questionnaire or questionnaires or interview or interviews or evaluation or evaluations or exam or exams or examination or examinations or test or tests or screening or screenings or assessment or assessments or index or indices or indexes or qualitative research or qualitative study or qualitative studies or checklist or checklists).tw.
6	or/4-5
7	adult/ or middle aged/ or young adult/ or aged/ or frail elderly/ or very elderly/
8	(adult or adults or young adult or young adults or middle age or middle aged or elderly or elder or oldest old or nonagenarian or nonagenarians or octogenarian or octogenarians or centenarian or centenarians or frail).tw.
9	or/7-8
10	delirium tremens/
11	(alcohol withdrawal syndrome or delirium tremens or alcohol withdrawal delirium).tw.
12	or/10-11
13	and/3,6,9
14	not/12-13
15	Limit to systematic reviews and meta-analysis
16	and/14-15

**CINAHL****Platform:** EBSCO, 1981-Present**Other Limits:** English language

S1	(MH "Delirium") OR (TI Delirium OR AB Delirium) OR (TI "Acute organic brain syndrome" OR AB "Acute organic brain syndrome") OR (TI "Acute confusion" OR AB "Acute confusion") OR (TI "Acute confusional state" OR AB "Acute confusional state") OR (TI "Acute brain syndrome" OR AB "Acute brain syndrome") OR (TI "Acute brain failure" OR AB "Acute brain failure") OR (TI "Acute brain dysfunction" OR AB "Acute brain dysfunction") OR (TI "Acute organic psychosyndrome" OR AB "Acute organic psychosyndrome") OR (TI "Acute organic psycho-syndrome" OR AB "Acute organic psycho-syndrome") OR (TI "Acute psycho-organic syndrome" OR AB "Acute psycho-organic syndrome") OR (TI "Acute psychoorganic syndrome" OR AB "Acute psychoorganic syndrome") OR (TI "Metabolic encephalopathy" OR AB "Metabolic encephalopathy") OR (TI "Clouded state" OR AB "Clouded state") OR (TI "Clouding of consciousness" OR AB "Clouding of consciousness")
S2	(MH "Checklists") OR (MH "Behavior Rating Scales") OR (MH "Interview Guides+") OR (MH "Psychological Tests+") OR (MH "Questionnaires+") OR (MH "Scales") OR (MH "Instrument Construction+") OR (MH "Patient Assessment") OR (MH "Health Screening") OR (MH "Neuropsychological Tests") OR (MH "Psychometrics") OR (MH "Interviews+") OR (MH "Clinical Assessment Tools") OR (MH "Short Portable Mental Status Questionnaire") OR MH "Qualitative studies" OR MH "Qualitative research" OR (TI scale OR AB scale) OR (TI scales OR AB scales) OR (TI instrument OR AB instrument) OR (TI instruments OR AB instruments) OR (TI measure OR AB measure) OR (TI measures OR AB measures) OR (TI questionnaire OR AB questionnaire) OR (TI questionnaires OR AB questionnaires) OR (TI interview OR AB interview) OR (TI interviews OR AB interviews) OR (TI evaluation OR AB evaluation) OR (TI evaluations OR AB evaluations) OR (TI examination OR AB examination) OR (TI examinations OR AB examinations) OR (TI exam OR AB exam) OR (TI exams OR AB exams) OR (TI test OR AB test) OR (TI tests OR AB tests) OR (TI screening OR AB screening) OR (TI screenings OR AB screenings) OR (TI assessment OR AB assessment) OR (TI assessments OR AB assessments) OR (TI index OR AB index) OR (TI indices OR AB indices) OR (TI indexes OR AB indexes) OR (TI checklist OR AB checklist) OR (TI checklists OR AB checklists)
S3	(MH "Adult+") OR (MH "Aged+") OR (TI Adult OR AB Adult) OR (TI Adults or AB Adults) OR (TI "Young adult" OR AB "Young adult") OR (TI "Young adults" OR AB "Young adults") OR (TI "Middle age" OR

	AB "Middle age") OR (TI "Middle aged" OR AB "Middle aged") OR (TI Elderly OR AB Elderly) OR (TI Elder OR AB Elder) OR (TI "Oldest old" OR AB "Oldest old") OR (TI Nonagenarian OR AB Nonagenarian) OR (TI Nonagenarians OR AB Nonagenarians) OR (TI Octogenarian OR AB Octogenarian) OR (TI Octogenarians OR AB Octogenarians) OR (TI Centenarian OR AB Centenarian) OR (TI Centenarians OR AB Centenarians) OR (TI Frail OR AB Frail)
S4	(MH "Alcohol Withdrawal Delirium") OR (TI "Delirium tremens" OR AB "Delirium tremens") OR (TI "Alcohol withdrawal delirium" OR AB "Alcohol withdrawal delirium")
S5	S1 AND S2 AND S3
S6	S5 NOT S4
S7	Limit to systematic reviews and meta-analysis
S8	S6 AND S7

### Web of Science Core Collection

**Platform:** Thomson Reuters, Science Citation Index Expanded, 1900-present; Social Sciences Citation Index, 1900-present; Arts & Humanities Citation Index, 1975-present; Conference Proceedings Citation Index-Science, 1990-present; Conference Proceedings Citation Index-Social Science & Humanities, 1990-present; Emerging Sources Citation Index, 2015-present

**Year Limits:** Publication Year 1974-Present

**Other Limits:** English language

#1	TS=(Delirium OR "Acute confusion" OR "Acute organic brain syndrome" OR "Acute confusional state" OR "Acute brain syndrome" OR "Acute brain failure" OR "Acute brain dysfunction" OR "Acute organic psychosyndrome" OR "Acute organic psycho-syndrome" OR "Acute psycho-organic syndrome" OR "Acute psychoorganic syndrome" OR "Metabolic encephalopathy" OR "Clouded state" OR "Clouding of consciousness")
#2	TS=(Scale OR Instrument OR Measure OR Questionnaire OR Interview OR Evaluation OR Examination OR Exam OR Test OR Screening OR Assessment OR Index OR Indices OR Indexes OR Checklist OR Tool OR "Qualitative study" OR "Qualitative studies")
#3	TS=(Adult OR "Young adult" OR "Young adults" OR "Middle age" OR "Middle aged" OR Elderly OR Elder OR "Oldest old" OR Nonagenarian OR Octogenarian OR Centenarian OR Frail)
#4	TS=("Alcohol withdrawal delirium" OR "Delirium tremens" OR "Alcohol withdrawal syndrome")
#5	#1 AND #2 AND #3
#6	#5 NOT #4
#7	Limit to systematic reviews and meta-analysis
#8	#6 AND #7

**PsycINFO****Platform:** EBSCO, 1880s-Present; ProQuest, 1806-Present**Year Limits:** 01/01/1974 to present**Other Limits:** English language; Adulthood (18 yrs & older)

S1	SU.EXACT("Delirium") OR TI(Delirium) OR AB(Delirium) OR TI("Acute confusion") OR AB("Acute confusion") OR TI("Acute organic brain syndrome") OR AB("Acute organic brain syndrome") OR TI("Acute confusional state") OR AB("Acute confusional state") OR TI("Acute brain syndrome") OR AB("Acute brain syndrome") OR TI("Acute brain failure") OR AB("Acute brain failure") OR TI("Acute brain dysfunction") OR AB("Acute brain dysfunction") OR TI("Acute organic psychosyndrome") OR AB("Acute organic psychosyndrome") OR TI("Acute organic psycho-syndrome") OR AB("Acute organic psycho-syndrome") OR TI("Acute psycho-organic syndrome") OR AB("Acute psycho-organic syndrome") OR TI("Acute psychoorganic syndrome") OR AB("Acute psychoorganic syndrome") OR TI("Metabolic encephalopathy") OR AB("Metabolic encephalopathy") OR TI("Clouded state") OR AB("Clouded state") OR TI("Clouding of consciousness") OR AB("Clouding of consciousness")
S2	SU.EXACT("Measurement") OR SU.EXACT("Achievement Measures") OR SU.EXACT("Aptitude Measures") OR SU.EXACT("Attitude Measurement") OR SU.EXACT("Attitude Measures") OR SU.EXACT("Body Sway Testing") OR SU.EXACT("Comprehension Tests") OR SU.EXACT("Creativity Measurement") OR SU.EXACT("Criterion Referenced Tests") OR SU.EXACT("Digit Span Testing") OR SU.EXACT("Group Testing") OR SU.EXACT("Individual Testing") OR SU.EXACT("Inventories") OR SU.EXACT("Multidimensional Scaling") OR SU.EXACT("Needs Assessment") OR SU.EXACT("Occupational Interest Measures") OR SU.EXACT("Perceptual Measures") OR SU.EXACT("Performance Tests") OR SU.EXACT("Preference Measures") OR SU.EXACT("Projective Testing Technique") OR SU.EXACT("Psychiatric Evaluation") OR SU.EXACT("Psychological Assessment") OR SU.EXACT("Psychometrics") OR SU.EXACT("Questionnaires") OR SU.EXACT("Rating Scales") OR SU.EXACT("Retention Measures") OR SU.EXACT("Screening") OR SU.EXACT("Screening Tests") OR SU.EXACT("Selection Tests") OR SU.EXACT("Sensorimotor Measures") OR SU.EXACT("Sociometric Tests") OR SU.EXACT("Speech and Hearing Measures") OR SU.EXACT("Standardized Tests") OR

SU.EXACT("Surveys") OR SU.EXACT("Symptom Checklists")  
 OR SU.EXACT("Testing") OR SU.EXACT("Verbal Tests") OR  
 SU.EXACT("Test Items") OR SU.EXACT("Content Analysis  
 (Test)") OR SU.EXACT("Difficulty Level (Test)") OR  
 SU.EXACT("Item Analysis (Test)") OR SU.EXACT("Item Content  
 (Test)") OR SU.EXACT("Item Response Theory") OR  
 SU.EXACT("Rating") OR SU.EXACT("Scaling (Testing)") OR  
 SU.EXACT("Scoring (Testing)") OR SU.EXACT("Test  
 Administration") OR SU.EXACT("Test Bias") OR  
 SU.EXACT("Test Forms") OR SU.EXACT("Test Interpretation")  
 OR SU.EXACT("Test Reliability") OR SU.EXACT("Test  
 Standardization") OR SU.EXACT("Test Validity") OR  
 SU.EXACT("Testing Methods") OR SU.EXACT("Adaptive  
 Testing") OR SU.EXACT("Cloze Testing") OR SU.EXACT("Essay  
 Testing") OR SU.EXACT("Forced Choice (Testing Method)") OR  
 SU.EXACT("Multiple Choice (Testing Method)") OR  
 SU.EXACT("Q Sort Testing Technique") OR  
 SU.EXACT("Questionnaires") OR SU.EXACT("General Health  
 Questionnaire") OR SU.EXACT("Interview Schedules") OR  
 SU.EXACT("Diagnostic Interview Schedule") OR  
 SU.EXACT("Structured Clinical Interview") OR  
 SU.EXACT("Interviews") OR SU.EXACT("Intake Interview") OR  
 SU.EXACT("Psychodiagnostic Interview") OR  
 SU.EXACT("Evaluation") OR SU.EXACT("Clinical Audits") OR  
 SU.EXACT("Geriatric Assessment") OR SU.EXACT("Psychiatric  
 Evaluation") OR SU.EXACT("Self Evaluation") OR  
 SU.EXACT("Screening") OR SU.EXACT("Health Screening") OR  
 SU.EXACT("Screening Tests") OR SU.EXACT("Psychological  
 Screening Inventory") OR SU.EXACT("Cognitive Assessment")  
 OR SU.EXACT("Neuropsychological assessment") OR  
 SU.EXACT("Behavioral Assessment") OR SU.EXACT("Likert  
 Scales") OR SU.EXACT("Test Construction") OR  
 SU.EXACT("Interviewing") OR SU.EXACT("Evaluation Criteria")  
 OR TI(scale) OR AB(scale) OR TI(scales) OR AB(scales) OR  
 TI(instrument) OR AB(instrument) OR TI(instruments) OR  
 AB(instruments) OR TI(measure) OR AB(measure) OR  
 TI(measures) OR AB(measures) OR TI(questionnaire) OR  
 AB(questionnaire) OR TI(questionnaires) OR AB(questionnaires)  
 OR TI(interview) OR AB(interview) OR TI(interviews) OR  
 AB(interviews) OR TI(evaluation) OR AB(evaluation) OR  
 TI(evaluations) OR AB(evaluations) OR TI(examination) OR  
 AB(examination) OR TI(examinations) OR AB(examinations) OR  
 TI(exam) OR AB(exam) OR TI(exams) OR AB(exams) OR  
 TI(test) OR AB(test) OR TI(tests) OR AB(tests) OR TI(screening)

	OR AB(screening) OR TI(screenings) OR AB(screenings) OR TI(assessment) OR AB(assessment) OR TI(assessments) OR AB(assessments) OR TI(index) OR AB(index) OR TI(indices) OR AB(indices) OR TI(indexes) OR AB(indexes) OR TI(checklist) OR AB(checklist) OR TI(checklists) OR AB(checklists)
S3	SU.EXACT("Alcohol Withdrawal") OR SU.EXACT("Drug Withdrawal") OR TI("Alcohol withdrawal delirium") OR AB("Alcohol withdrawal delirium") OR SU.EXACT("Delirium Tremens") OR TI("Delirium tremens") OR AB("Delirium tremens")
S4	S1 AND S2
S5	S4 NOT S3
S6	Limit to systematic reviews and meta-analysis
S7	S5 AND S6



**Cochrane Library****Platform:** Wiley Online Library**Year Limits:** Publication year 1974-present**Other Limits:** Cochrane Reviews, Database of Abstracts of Reviews of Effects (DARE), Cochrane Central Register of Controlled Trials (CENTRAL)

#1	[mh Delirium] or "Delirium":ti,ab or "Acute confusion":ti,ab or "Acute organic brain syndrome":ti,ab or "Acute confusional state":ti,ab or "Acute brain syndrome":ti,ab or "Acute brain failure":ti,ab "Acute brain dysfunction":ti,ab or "Acute organic psychosyndrome":ti,ab or "Acute organic psycho-syndrome":ti,ab or "Acute psycho-organic syndrome":ti,ab or "Acute psychoorganic syndrome":ti,ab or "Metabolic encephalopathy":ti,ab or "Clouded state":ti,ab or "Clouding of consciousness":ti,ab
#2	[mh "Psychiatric Status Rating Scales"] or [mh "Neuropsychological Tests"] or [mh Psychometrics] or [mh "Mass Screening"] or [mh "Geriatric Assessment"] or [mh "Psychological Tests"] or [mh "Surveys and Questionnaires"] or [mh "Interview, Psychological"] or [mh "Qualitative Research"] or [mh "Mental Status Schedule"] or [mh Checklist] or "Scale":ti,ab or "Scales":ti,ab or "Instrument":ti,ab or "Instruments":ti,ab or "Measure":ti,ab or "Measures":ti,ab or "Questionnaire":ti,ab or "Questionnaires":ti,ab or "Interview":ti,ab or "Interviews":ti,ab or "Evaluation":ti,ab or "Evaluations":ti,ab or "Exam":ti,ab or "Exams":ti,ab or "Examination":ti,ab or "Examinations":ti,ab or "Test":ti,ab or "Tests":ti,ab or "Screening":ti,ab or "Screenings":ti,ab or "Assessment":ti,ab or "Assessments":ti,ab or "Index":ti,ab or "Indices":ti,ab or "Indexes":ti,ab or "Qualitative Research":ti,ab or "Qualitative Study":ti,ab or "Qualitative Studies":ti,ab or "Checklist":ti,ab or "Checklists":ti,ab
#3	[mh Adult] or [mh "Young Adult"] or [mh Aged] or [mh "Aged, 80 and over"] or [mh "Frail Elderly"] or "Adults":ti,ab or "Adult":ti,ab or "Young Adults":ti,ab or "Young Adult":ti,ab or "Middle age":ti,ab or "Middle aged":ti,ab or "Elderly":ti,ab or "Elder":ti,ab or "Oldest Old":ti,ab or "Nonagenarian":ti,ab or "Nonagenarians":ti,ab or "Octogenarian":ti,ab or "Octogenarians":ti,ab or "Centenarian":ti,ab or "Centenarians":ti,ab or "Frail":ti,ab
#4	[mh "Alcohol withdrawal delirium"] or "Alcohol withdrawal syndrome":ti,ab or "Delirium tremens":ti,ab or "Alcohol withdrawal delirium":ti,ab
#5	#1 and #2 and #3
#6	#5 not #4
#7	Limit to systematic reviews and meta-analysis
#8	#6 and #7

**Table 2.7. COSMIN-guided psychometric review (adapted from RN Jones 2019)**

**Definitions and scoring approach:**

<b>Criterion</b>	<b>Definition</b>
Effect indicators	Effect indicators are influenced by or related to delirium, such as signs and symptoms of delirium. Effect indicators are appropriate for use in a measurement instrument. Cause or formative indicators are factors that might cause delirium (e.g., signs of infection), and would not be appropriate to include. Studies were given a score of 1 if all items were effect indicators or a score of 0 if the items included potential causative factors.
Content validity	Content validity refers to ensuring that all items capture relevant aspects of delirium. For instance, this can be assessed by face validity reviews involving experts, literature reviews, etc. If the study mentioned assessing content validity, then it was scored 1 otherwise failure to mention was scored 0.
Internal consistency	Internal consistency refers to how each item relates to the others in the instrument. It is important to make sure the instrument assesses a single underlying construct, delirium identification. If the authors report internal consistency reliability with a value such as Cronbach's coefficient alpha or McDonald's omega coefficient, then a point was awarded. However, if a sample size of less than 50 was used in calculating internal consistency they lose ½ point. If the authors failed to mention assessment of internal consistency, they were awarded no points.
Inter-rater reliability	Refers to assessments of the agreement between two or more raters when making ratings on a single patient or research participant. We recorded any mention and statistics given including Pearson correlation coefficient, intra-class correlation coefficient, or Kappa statistics. If the authors mentioned inter-rater reliability, they were given a point and deducted a half point for using a sample size less than 50. They were given no points if they failed to mention any assessment of inter-rater reliability.
Construct validity	Describes how well and instrument measuring a construct correlates with other instruments measuring the same construct, in this case delirium identification. If this comparison

	was performed, we recorded any correlation coefficients and awarded a point. We deducted a half point for using a sample size less than 50. They were given no points if they failed to mention any comparison instruments.
External (or criterion-related) validity	refers to comparison of the proposed instrument against a reference standard used for delirium case identification. We recorded the reference standard and awarded a point if assessed. We deducted a half point for using a sample size less than 50. They were given no points if they failed to mention any reference standard.

### **References**

1. Jones RN, Cizginer S, Pavlech L, et al. Assessment of instruments for measurement of delirium severity: a systematic review. *JAMA internal medicine*. 2019;179(2):231-239.
2. Terwee CB, B ML, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research*. 2012;4(1573-2649 (Electronic)):651-657.
3. Mokkink L, Terwee C, Patrick D, et al. COSMIN checklist manual. Amsterdam: COSMIN network; 2012.
4. De Vet HC, Terwee CB, Mokkink LB, Knol DL. *Measurement in medicine: a practical guide*. Cambridge University Press; 2011.
5. Mokkink LB, Terwee CB, Patrick DL, et al. COSMIN checklist manual. Amsterdam: University Medical Center. 2012.

6. Mokkink LB, Terwee CB, Stratford PW, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research*. 2009;18(3):313-333.

**Table 2.8. List of citations of eligible articles**

1. Adamis D, Sharma N, Whelan PJP, Macdonald AJD. Delirium scales: a review of current evidence. *Aging & Mental Health*. 2010;14(5):543-55. doi: 10.1080/13607860903421011. PubMed PMID: 105049095.
2. Barr J, Fraser GL, Puntillo K, Ely EW, Gélinas C, Dasta JF, et al. Clinical practice guidelines for the management of pain, agitation, and delirium in adult patients in the intensive care unit. *Critical care medicine*. 2013;41(1):263-306.
3. Barr J, Kishman Jr CP, Jaeschke R. The methodological approach used to develop the 2013 pain, agitation, and delirium clinical practice guidelines for adult ICU patients. *Critical Care Medicine*. 2013:S1-S15. doi: 10.1097/CCM.0b013e3182a167d7. PubMed PMID: 107970311.
4. Barron EA, Holmes J. Delirium within the emergency care setting, occurrence and detection: a systematic review. *Emergency Medicine Journal*. 2013;30(4):263-8. doi: 10.1136/emmermed-2011-200586. PubMed PMID: WOS:000316314200003.
5. Beales LK, Mercuri M. BET 1: Screening for delirium within the emergency department. *Emergency medicine journal : EMJ*. 2016;33(10):741-3. Epub 2016/09/22. doi: 10.1136/emmermed-2016-206204.1. PubMed PMID: 27651501.
6. Bhat R, Rockwood K. Inter-rater reliability of delirium rating scales. *Neuroepidemiology*. 2005;25(1):48-52. Epub 2005/04/28. doi: 10.1159/000085440. PubMed PMID: 15855805.
7. Brooks PB. Postoperative Delirium in Elderly Patients. *AJN American Journal of Nursing*. 2012;112(9):38-51. PubMed PMID: 108151543.
8. Bruce AJ, Ritchie CW, Blizard R, Lai R, Raven P. The incidence of delirium associated with orthopedic surgery: A meta-analytic review. *International Psychogeriatrics*. 2007;19(2):197-214. doi: 10.1017/S104161020600425X.
9. Brummel NE, Vasilevskis EE, Han JH, Boehm L, Pun BT, Ely EW. Implementing delirium screening in the ICU: secrets to success. *Critical care medicine*. 2013;41(9):2196-208. Epub 2013/07/31. doi: 10.1097/CCM.0b013e31829a6f1e. PubMed PMID: 23896832; PubMed Central PMCID: PMC3772682.
10. Carin-Levy G, Mead GE, Nicol K, Rush R, van Wijck F. Delirium in acute stroke: screening tools, incidence rates and predictors: a systematic review. *Journal of Neurology*. 2012;259(8):1590-9. doi: 10.1007/s00415-011-6383-4. PubMed PMID: WOS:000307267300010.
11. Cull E, Phillips NM, Kent B, Mistarz R. Risk factors for incident delirium in acute medical in-patients. A systematic review. *JBIC Database of Systematic Reviews and Implementation Reports*. 2012;10:S51-S62.
12. de Rooij SE, Schuurmans MJ, van der Mast RC, Levi M. Clinical subtypes of delirium and their relevance for daily clinical practice: A systematic review. *International Journal of Geriatric Psychiatry*. 2005;20(7):609-15. doi: 10.1002/gps.1343.
13. Devlin JW, Fong JJ, Fraser GL, Riker RR. Delirium assessment in the critically ill. *Intensive care medicine*. 2007;33(6):929-40. Epub 2007/04/03. doi: 10.1007/s00134-007-0603-5. PubMed PMID: 17401550.

14. El Hussein M, Hirst S, Salyers V. Factors that contribute to underrecognition of delirium by registered nurses in acute care settings: a scoping review of the literature to explain this phenomenon. *Journal of clinical nursing*. 2015;24(7-8):906-15.
15. Fan YY, Guo Y, Li QJ, Zhu XM. A Review: Nursing of Intensive Care Unit Delirium. *Journal of Neuroscience Nursing*. 2012;44(6):307-16. doi: 10.1097/JNN.0b013e3182682f7f. PubMed PMID: WOS:000310756000004.
16. Fick DM, Agostini JV, Inouye SK. Delirium superimposed on dementia: a systematic review. *Journal of the American Geriatrics Society*. 2002;50(10):1723-32. doi: 10.1046/j.1532-5415.2002.50468.x. PubMed PMID: 106691685.
17. Flaherty JH, Yue J, Rudolph JL. Dissecting Delirium: Phenotypes, Consequences, Screening, Diagnosis, Prevention, Treatment, and Program Implementation. *Clinics in geriatric medicine*. 2017;33(3):393-413. Epub 2017/07/12. doi: 10.1016/j.cger.2017.03.004. PubMed PMID: 28689571.
18. Fong TG, Tulebaev SR, Inouye SK. Delirium in elderly adults: diagnosis, prevention and treatment. *Nature reviews Neurology*. 2009;5(4):210-20. Epub 2009/04/07. doi: 10.1038/nrneurol.2009.24. PubMed PMID: 19347026; PubMed Central PMCID: PMC3065676.
19. Forsberg MM. Delirium Update for Postacute Care and Long-Term Care Settings: A Narrative Review. *The Journal of the American Osteopathic Association*. 2017;117(1):32-8. Epub 2017/01/06. doi: 10.7556/jaoa.2017.005. PubMed PMID: 28055085.
20. Gelinac C, Berube M, Chevrier A, Pun BT, Ely EW, Skrobik Y, et al. Delirium Assessment Tools for Use in Critically Ill Adults: A Psychometric Analysis and Systematic Review. *Critical care nurse*. 2018;38(1):38-49. Epub 2018/02/14. doi: 10.4037/ccn2018633. PubMed PMID: 29437077.
21. Girard TD, Pandharipande PP, Ely EW. Delirium in the intensive care unit. *Critical Care*. 2008;12. doi: 10.1186/cc6149. PubMed PMID: WOS:000257633700003.
22. González-Gil T. Interventions for preventing delirium in older people in institutional long-term care. *International journal of nursing studies*. 2016;55:133-4.
23. Greer N, Rossom R, Anderson P, MacDonald R, Tacklind J, Rutks I, et al. VA Evidence-based Synthesis Program Reports. Delirium: Screening, Prevention, and Diagnosis - A Systematic Review of the Evidence. Washington (DC): Department of Veterans Affairs (US); 2011.
24. Gusmao-Flores D, Figueira Salluh JI, Chalhub RT, Quarantini LC. The confusion assessment method for the intensive care unit (CAM-ICU) and intensive care delirium screening checklist (ICDSC) for the diagnosis of delirium: a systematic review and meta-analysis of clinical studies. *Critical Care*. 2012;16(4). doi: 10.1186/cc11407.
25. Halloway S. A family approach to delirium: a review of the literature. *Aging & Mental Health*. 2014;18(2):129-39. doi: 10.1080/13607863.2013.814102. PubMed PMID: 104015458.
26. Hendry K, Hill E, Quinn TJ, Evans J, Stott DJ. Single screening questions for cognitive impairment in older people: A systematic review. *Age and Ageing*. 2015;44(2):322-6. doi: 10.1093/ageing/afu167.

27. Hosie A, Davidson PM, Agar M, Sanderson CR, Phillips J. Delirium prevalence, incidence, and implications for screening in specialist palliative care inpatient settings: A systematic review. *Palliative Medicine*. 2013;27(6):486-98. doi: 10.1177/0269216312457214. PubMed PMID: 104292120.
28. Hsieh SJ, Shum M, Lee A, Al-Othman F, Gong MN. Cigarette smoking as a risk factor for delirium in hospitalized patients-a systematic review. *American Journal of Respiratory and Critical Care Medicine*. 2012;185.
29. Inouye SK, Westendorp RGJ, Saczynski JS. Delirium in elderly people. *The Lancet*. 2014;383(9920):911-22. doi: 10.1016/S0140-6736(13)60688-1.
30. Jayita D, Wand ARF. Delirium Screening: A Systematic Review of Delirium Screening Tools in Hospitalized Patients. *Gerontologist*. 2015;55(6):1079-99. doi: 10.1093/geront/gnv100. PubMed PMID: 111186039.
31. Khan BA, Zawahiri M, Campbell NL, Fox GC, Weinstein EJ, Nazir A, et al. Delirium in hospitalized patients: Implications of current evidence on clinical practice and future avenues for research-A systematic evidence review. *Journal of Hospital Medicine*. 2012;7(7):580-9. doi: 10.1002/jhm.1949.
32. LaMantia MA, Messina FC, Hobgood CD, Miller DK. Screening for delirium in the emergency department: a systematic review. *Annals of Emergency Medicine*. 2014;63(5):551-60.e2. doi: 10.1016/j.annemergmed.2013.11.010. PubMed PMID: 103950119.
33. Lawlor PG, Bush SH. Delirium diagnosis, screening and management. *Current Opinion in Supportive and Palliative Care*. 2014;8(3):286-95. doi: 10.1097/SPC.0000000000000062. PubMed PMID: WOS:000340512300017.
34. Leonard MM, Nikolaichuk C, Meagher DJ, Barnes C, Gaudreau J-D, Watanabe S, et al. Practical assessment of delirium in palliative care. *Journal of Pain and Symptom Management*. 2014;48(2):176-90. doi: 10.1016/j.jpainsymman.2013.10.024. PubMed PMID: 2014-34023-007.
35. Levkoff S, Liptzin B, Cleary P, Reilly CH, Evans D. Review of research instruments and techniques used to detect delirium. *International psychogeriatrics*. 1991;3(2):253-71. Epub 1991/01/01. PubMed PMID: 1811778.
36. Lindroth H, Bratzke L, Purvis S, Brown R, Coburn M, Mrkobrada M, et al. Systematic review of prediction models for delirium in the older adult inpatient. *BMJ open*. 2018;8(4):e019223. Epub 2018/05/01. doi: 10.1136/bmjopen-2017-019223. PubMed PMID: 29705752; PubMed Central PMCID: PMC5931306.
37. Maldonado JR. Delirium in the acute care setting: characteristics, diagnosis and treatment. *Critical care clinics*. 2008;24(4):657-722, vii. Epub 2008/10/22. doi: 10.1016/j.ccc.2008.05.008. PubMed PMID: 18929939.
38. Marshall MC, Soucy MD. Delirium in the intensive care unit. *Critical care nursing quarterly*. 2003;26(3):172-8. Epub 2003/08/22. PubMed PMID: 12930032.
39. Martocchia A, Curto M, Comite F, Scaccianoce S, Girardi P, Ferracuti S, et al. The Prevention and Treatment of Delirium in Elderly Patients Following Hip Fracture Surgery. *Recent patents on CNS drug discovery*. 2015;10(1):55-64. Epub 2015/02/18. PubMed PMID: 25687439.

40. Mattoo SK, Grover S, Gupta N. Delirium in general practice. *The Indian journal of medical research*. 2010;131:387-98. Epub 2010/04/27. PubMed PMID: 20418552.
41. Milisen K, Foreman MD, Godderis J, Abraham IL, Broos PL. Delirium in the hospitalized elderly: nursing assessment and management. *The Nursing clinics of North America*. 1998;33(3):417-39. Epub 1998/08/28. PubMed PMID: 9719689.
42. Mitchell AJ, Shukla D, Ajumal HA, Stubbs B, Tahir TA. The Mini-Mental State Examination as a diagnostic and screening test for delirium: systematic review and meta-analysis. *General Hospital Psychiatry*. 2014;36(6):627-33. doi: 10.1016/j.genhosppsych.2014.09.003. PubMed PMID: 103920167.
43. Moraga AV, Rodriguez-Pascual C. Accurate diagnosis of delirium in elderly patients. *Current opinion in psychiatry*. 2007;20(3):262-7. Epub 2007/04/07. doi: 10.1097/YCO.0b013e3280ec52e5. PubMed PMID: 17415080.
44. Morandi A, McCurley J, Vasilevskis EE, Fick DM, Bellelli G, Lee P, et al. Tools to detect delirium superimposed on dementia: A systematic review. *Journal of the American Geriatrics Society*. 2012;60(11):2005-13. PubMed PMID: 2012-30847-001.
45. Nazemi AK, Gowd AK, Carmouche JJ, Kates SL, Albert TJ, Behrend CJ. Prevention and Management of Postoperative Delirium in Elderly Patients Following Elective Spinal Surgery. *Clinical Spine Surgery*. 2017;30(3):112-9. doi: 10.1097/BSD.0000000000000467.
46. Neufeld KJ, Nelliot A, Inouye SK, Ely EW, Bienvenu OJ, Lee HB, et al. Delirium diagnosis methodology used in research: A survey-based study. *The American Journal of Geriatric Psychiatry*. 2014;22(12):1513-21. doi: 10.1016/j.jagp.2014.03.003. PubMed PMID: 2014-48897-020.
47. Nitchingham A, Kumar V, Shenkin S, Ferguson KJ, Caplan GA. A systematic review of neuroimaging in delirium: Predictors, correlates and consequences. *International Journal of Geriatric Psychiatry*. 2017. doi: 10.1002/gps.4724.
48. Oh ES, Fong TG, Hshieh TT, Inouye SK. Delirium in Older Persons: Advances in Diagnosis and Treatment. *JAMA: Journal of the American Medical Association*. 2017;318(12):1161-74. doi: 10.1001/jama.2017.12067. PubMed PMID: 125412347.
49. Popeo DM. Delirium in older adults. *The Mount Sinai journal of medicine, New York*. 2011;78(4):571-82. Epub 2011/07/13. doi: 10.1002/msj.20267. PubMed PMID: 21748745; PubMed Central PMCID: PMC3136888.
50. Popp J, Arlt S. Prevention and treatment options for postoperative delirium in the elderly. *Current Opinion in Psychiatry*. 2012;25(6):515-21. doi: 10.1097/YCO.0b013e328357f51c.
51. Quispel-Aggenbach DWP, Holtman GA, Zwartjes H, Zuidema SU, Luijendijk HJ. Attention, arousal and other rapid bedside screening instruments for delirium in older patients: a systematic review of test accuracy studies. *Age and ageing*. 2018. Epub 2018/04/27. doi: 10.1093/ageing/afy058. PubMed PMID: 29697753.
52. Robertsson B. Assessment scales in delirium. *Dementia and geriatric cognitive disorders*. 1999;10(5):368-79. Epub 1999/09/04. doi: 10.1159/000017173. PubMed PMID: 10473942.
53. Rood P, Huisman-de Waal G, Vermeulen H, Schoonhoven L, Pickkers P, van den Boogaard M. Effect of organisational factors on the variation in incidence of delirium



- in intensive care unit patients: A systematic review and meta-regression analysis. *Australian critical care : official journal of the Confederation of Australian Critical Care Nurses*. 2018;31(3):180-7. Epub 2018/03/17. doi: 10.1016/j.aucc.2018.02.002. PubMed PMID: 29545081.
54. Rosen T, Connors S, Clark S, Halpern A, Stern ME, DeWald J, et al. Assessment and Management of Delirium in Older Adults in the Emergency Department: Literature Review to Inform Development of a Novel Clinical Protocol. *Advanced emergency nursing journal*. 2015;37(3):183-96; quiz E3. Epub 2015/07/29. doi: 10.1097/tme.000000000000066. PubMed PMID: 26218485; PubMed Central PMCID: PMC4633298.
55. Rosgen B, Krewulak K, Demiantschuk D, Ely EW, Davidson JE, Stelfox HT, et al. Validation of Caregiver-Centered Delirium Detection Tools: A Systematic Review. *Journal of the American Geriatrics Society*. 2018;66(6):1218-25. doi: 10.1111/jgs.15362. PubMed PMID: 130769850.
56. Salluh JIF, Wang H, Schneider EB, Nagaraja N, Yenokyan G, Damluji A, et al. Outcome of delirium in critically ill patients: Systematic review and meta-analysis. *BMJ (Online)*. 2015;350:1-10. doi: 10.1136/bmj.h2538.
57. Saxena S, Lawley D. Delirium in the elderly: a clinical review. *Postgraduate medical journal*. 2009;85(1006):405-13. Epub 2009/07/28. doi: 10.1136/pgmj.2008.072025. PubMed PMID: 19633006.
58. Schievelde JNM, van Zwieten JJ. From Pediatrics to Geriatrics: Toward a Unified Standardized Screening Tool for Delirium: A Thought Experiment. *Critical Care Medicine*. 2016;44(9):1778-80. doi: 10.1097/CCM.0000000000001485. PubMed PMID: 117500439.
59. Schuurmans MJ, Deschamps PI, Markham SW, Shortridge-Baggett LM, Duursma SA. The measurement of delirium: review of scales. *Research and theory for nursing practice*. 2003;17(3):207-24. Epub 2003/12/06. PubMed PMID: 14655974.
60. Schuurmans MJ, Duursma SA, Shortridge-Baggett LM. Early recognition of delirium: review of the literature. *Journal of clinical nursing*. 2001;10(6):721-9. Epub 2002/02/02. PubMed PMID: 11822843.
61. Shi Q, Warren L, Saposnik G, MacDermid JC. Confusion assessment method: A systematic review and meta-analysis of diagnostic accuracy. *Neuropsychiatric Disease and Treatment*. 2013;9. PubMed PMID: 2013-34061-001.
62. Smith T, Hameed Y, Cross J, Sahota O, Fox C. Assessment of people with cognitive impairment and hip fracture: A systematic review and meta-analysis. *Archives of Gerontology and Geriatrics*. 2013;57(2):117-26. doi: 10.1016/j.archger.2013.04.009.
63. Suwanpasu S, Sattayasomboon Y. Hyperactive and hypoactive psychomotor subtypes of delirium in demented and nondemented elderly patients with hip fractures: Systematic review and meta-analysis. *Asian Biomedicine*. 2015;9(4):441-53. doi: 10.5372/1905-7415.0904.413.
64. Tamune H, Yasugi D. How can we identify patients with delirium in the emergency department?: A review of available screening and diagnostic tools. *The American journal of emergency medicine*. 2017;35(9):1332-4. Epub 2017/06/03. doi: 10.1016/j.ajem.2017.05.026. PubMed PMID: 28571901.

65. Trzepacz PT. Delirium - Advances in diagnosis, pathophysiology, and treatment. *Psychiatric Clinics of North America*. 1996;19(3):429-+. doi: 10.1016/S0193-953X(05)70299-9. PubMed PMID: WOS:A1996VB57700003.
66. van Velthuisen EL, Zwakhalen SM, Warnier RM, Mulder WJ, Verhey FR, Kempen GI. Psychometric properties and feasibility of instruments for the detection of delirium in older hospitalized patients: a systematic review. *International journal of geriatric psychiatry*. 2016;31(9):974-89. Epub 2016/02/24. doi: 10.1002/gps.4441. PubMed PMID: 26898375.
67. Wei LA, Fearing MA, Sternberg EJ, Inouye SK. The Confusion Assessment Method: A Systematic Review of Current Usage. *Journal of the American Geriatrics Society*. 2008;56(5):823-30. doi: 10.1111/j.1532-5415.2008.01674.x. PubMed PMID: 105736606.
68. Weinrich S, Sarna L. Delirium in the older person with cancer. *Cancer*. 1994;74(7 Suppl):2079-91. Epub 1994/10/01. PubMed PMID: 8087775.
69. Wong CL, Holroyd-Leduc J, Simel DL, Straus SE. Does this patient have delirium? Value of bedside instruments. *JAMA: Journal of the American Medical Association*. 2010;304(7):779-86. doi: 10.1001/jama.2010.1182. PubMed PMID: 2010-18198-002.
70. Mulkey MA, Roberson DW, Everhart DE, Hardin SR. Choosing the Right Delirium Assessment Tool. *Journal of Neuroscience Nursing*. 2018;50(6):343-8. doi: 10.1097/jnn.000000000000403. PubMed PMID: WOS:000450434500006
71. Dylan F, Byrne G, Mudge AM. Delirium risk in non-surgical patients: systematic review of predictive tools. *Arch Gerontol Geriatr*. 2019;83:292-302. Epub 2019/05/29. doi: 10.1016/j.archger.2019.05.013. PubMed PMID: 31136886.
72. Oh-Park M, Chen P, Romel-Nichols V, Hreha K, Boukrina O, Barrett AM. Delirium Screening and Management in Inpatient Rehabilitation Facilities. *Am J Phys Med Rehabil*. 2018;97(10):754-62. Epub 2018/05/10. doi: 10.1097/phm.0000000000000962. PubMed PMID: 29742533; PMCID: PMC6148375.
73. Balkova M, Tomagova M. Use of measurement tools for screening of postoperative delirium in nursing practice. *Central European Journal of Nursing and Midwifery*. 2018;9(3):897-904. doi: <http://dx.doi.org/10.15452/CEJNM.2018.09.0021>.
74. Sevcikova B, Kubsova HM, Satekova L, Gurkova E. Delirium screening instruments administered by nurses for hospitalized patients - Literature review. *Central European Journal of Nursing and Midwifery*. 2019;10(4):1167-78. doi: <http://dx.doi.org/10.15452/CEJNM.2019.10.0028>.
75. Stokholm J, Steenholt JV, Csilag C, Kjaer TW, Christensen T. Delirium Assessment in Acute Stroke: A Systematic Review and Meta-Analysis of Incidence, Assessment Tools, and Assessment Frequencies. *J Cent Nerv Syst Dis*. 2019;11:1179573519897083. Epub 2020/01/08. doi: 10.1177/1179573519897083. PubMed PMID: 31908562; PMCID: PMC6937530.

**Table 2.9. List of excluded instruments with reasons**

Instruments removed because severity only: 8

3-Minute Diagnostic Assessment-Severity (3D-CAM-S)  
 Communication Capacity Scale and Agitation Distress Scale (CCS-ADS)  
 Confusion Assessment Method Severity Score (CAM-S)  
 Confusion State Evaluation (CSE)  
 Delirium Assessment Scale (DAS)  
 Delirium Index (DI)  
 Delirium-O-Meter  
 Delirium Severity Scale (DSS)

Instruments removed because ICU only: 5

Cognitive Test for Delirium (CTD)  
 Confusion Assessment Method for the Intensive Care Unit (CAM-ICU)  
 Confusion Assessment Method for the Intensive Care Unit-7 (CAM-ICU-7)  
 Delirium detection Score (DDS)  
 Intensive Care Delirium Screening Checklist (ICDSC)

Instruments removed because subtype (hypoactive or hyperactive) only: 2

Delirium Motoric Checklist  
 Motor Subtype Scale

Instruments removed because used to define risk for developing delirium: 1

Delirium Elderly At Risk Instrument (DEAR)

Instruments removed because published before 1974: 1

Delirium Scale (D-Scale)

Instruments removed because a test of attention only and not delirium identification: 1

DelApp

Instruments removed because case study only: 1

Delirium in Cancer Patients

**Table 2.10. List of citations of all instruments included**

1. Albert MS, Levkoff SE, Reilly C, Liptzin B, Pilgrim D, Cleary PD, et al. The delirium symptom interview: an interview for the detection of delirium symptoms in hospitalized patients. *Topics in geriatrics*. 1992;5(1):14-21.
2. Bellelli G, Morandi A, Davis DHJ, Mazzola P, Turco R, Gentile S, et al. Validation of the 4AT, a new instrument for rapid delirium screening: a study in 234 hospitalised older people. *Age and ageing*. 2014;43(4):496-502.
3. Breitbart W, Rosenfeld B, Roth A, Smith MJ, Cohen K, Passik S. The memorial delirium assessment scale. *Journal of pain and symptom management*. 1997;13(3):128-37.
4. Cacchione PZ. Four acute confusion assessment instruments: reliability and validity for use in long-term care facilities. *Journal of Gerontological Nursing*. 2002;28(1):12-9.
5. Dosa D, Intrator O, McNicoll L, Cang Y, Teno J. Preliminary derivation of a nursing home confusion assessment method based on data from the minimum data set. *Journal of the American Geriatrics Society*. 2007;55(7):1099-105.
6. Funk SG, Tornquist EM, Champagne MT, Wiese R. *Key aspects of elder care: Managing falls, incontinence, and cognitive impairment*: Springer Publishing Company; 1992.
7. Gaudreau J-D, Gagnon P, Harel F, Roy M-A. Impact on delirium detection of using a sensitive instrument integrated into clinical practice. *General hospital psychiatry*. 2005;27(3):194-9.
8. Grossmann FF, Hasemann W, Graber A, Bingisser R, Kressig RW, Nickel CH. Screening, detection and management of delirium in the emergency department—a pilot study on the feasibility of a new algorithm for use in older emergency department patients: the modified Confusion Assessment Method for the Emergency Department (mCAM-ED). *Scandinavian journal of trauma, resuscitation and emergency medicine*. 2014;22(1):19.
9. Gustafson L, editor *Organic Brain Syndrome Scale (OBSscale)*. A new rating scale for evaluation of confusional states and other organic brain syndromes 1985 1985.
10. Han JH, Wilson A, Vasilevskis EE, Shintani A, Schnelle JF, Dittus RS, et al. Diagnosing delirium in older emergency department patients: validity and reliability of the delirium triage screen and the brief confusion assessment method. *Annals of emergency medicine*. 2013;62(5):457-65.
11. Inouye SK, Leo-Summers L, Zhang Y, Bogardus Jr ST, Leslie DL, Agostini JV. A chart-based method for identification of delirium: validation compared with interviewer ratings using the confusion assessment method. *Journal of the American Geriatrics Society*. 2005;53(2):312-8.
12. Inouye SK, van Dyck CH, Alessi CA, Balkin S, Siegal AP, Horwitz RI. Clarifying confusion: the confusion assessment method: a new method for detection of delirium. *Annals of internal medicine*. 1990;113(12):941-8.

13. Kean J, Trzepacz PT, Murray LL, Abell M, Trexler L. Initial validation of a brief provisional diagnostic scale for delirium. *Brain injury*. 2010;24(10):1222-30.
14. Lewis LM, Miller DK, Morley JE, Nork MJ, Lasater LC. Unrecognized delirium in ED geriatric patients. *The American journal of emergency medicine*. 1995;13(2):142-5.
15. Lin HS, Eeles E, Pandey S, Pinsker D, Brasch C, Yerkovich S. Screening in delirium: A pilot study of two screening tools, the Simple Query for Easy Evaluation of Consciousness and Simple Question in Delirium. *Australasian journal on ageing*. 2015;34(4):259-64.
16. Marcantonio ER, Ngo LH, O'Connor M, Jones RN, Crane PK, Metzger ED, et al. 3D-CAM: derivation and validation of a 3-minute diagnostic interview for CAM-defined delirium: a cross-sectional diagnostic test study. *Annals of internal medicine*. 2014;161(8):554-61.
17. Miller PS, Richardson JS, Jyu CA, Lemay JS, Hiscock M, Keegan DL. Association of low serum anticholinergic levels and cognitive impairment in elderly presurgical patients. *Am J Psychiatry*. 1988;145:342-5.
18. Neelon VJ, Champagne MT, Carlson JR, Funk SG. The NEECHAM Confusion Scale: construction, validation, and clinical testing. *Nursing research*. 1996;45(6):324-30.
19. Rhodius-Meester HFM, van Campen J, Fung W, Meagher DJ, van Munster BC, de Jonghe JFM. Development and validation of the informant assessment of geriatric delirium scale (I-AGeD). *Recognition of delirium in geriatric patients. European Geriatric Medicine*. 2013;4(2):73-7.
20. Salih SA, Paul S, Klein K, Lakhan P, Gray L. Screening for delirium within the interRAI acute care assessment system. *The journal of nutrition, health & aging*. 2012;16(8):695-700.
21. Sands MB, Dantoc BP, Hartshorn A, Ryan CJ, Lujic S. Single question in delirium (SQiD): testing its efficacy against psychiatrist interview, the confusion assessment method and the memorial delirium assessment scale. *Palliative Medicine*. 2010;24(6):561-5.
22. Schuurmans MJ, Shortridge-Baggett LM, Duursma SA. The Delirium Observation Screening Scale: a screening instrument for delirium. *Research and theory for nursing practice*. 2003;17(1):31-50.
23. Steis MR, Evans L, Hirschman KB, Hanlon A, Fick DM, Flanagan N, et al. Screening for delirium using family caregivers: Convergent validity of the Family Confusion Assessment Method and interviewer-rated Confusion Assessment Method. *Journal of the American Geriatrics Society*. 2012;60(11):2121-6.
24. Stillman MJ, Rybicki LA. The bedside confusion scale: development of a portable bedside test for confusion and its application to the palliative medicine population. *Journal of palliative medicine*. 2000;3(4):449-56.
25. Treloar AJ, Macdonald AJD. OUTCOME OF DELIRIUM: PART 1. Outcome of Delirium Diagnosed by DSM-III-R, ICD-10 and CAMDEX and Derivation of the Reversible Cognitive Dysfunction Scale Among Acute Geriatric Inpatients. *International journal of geriatric psychiatry*. 1997;12(6):609-13.

26. Trzepacz PT, Baker RW, Greenhouse J. A symptom rating scale for delirium. *Psychiatry research*. 1988;23(1):89-97.
27. Trzepacz PT, Mittal D, Torres R, Kanary K, Norton J, Jimerson N. Validation of the Delirium Rating Scale-revised-98: comparison with the delirium rating scale and the cognitive test for delirium. *The Journal of neuropsychiatry and clinical neurosciences*. 2001;13(2):229-42.
28. Vermeersch PEH. The clinical assessment of confusion-A. *Applied Nursing Research*. 1990;3(3):128-33.
29. Voyer P, Champoux N, Desrosiers J, Landreville P, McCusker J, Monette J, et al. Recognizing acute delirium as part of your routine [RADAR]: a validation study. *BMC nursing*. 2015;14(1):19.
30. Williams MA. Delirium/acute confusional states: evaluation devices in nursing. *International Psychogeriatrics*. 1991;3(2):301-8.

## ***Acknowledgment***

### Published Manuscript Co-Author Contributions

Mr. Helfand conceived of the project, collected the search, organized and convened the expert panel, synthesized expert panel feedback, did all analyses, wrote the manuscript, created all tables and figures. Mr. Helfand and Dr. Jones had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Other contributions include:

Concept and design: Helfand, Jones, Inouye, Boudreaux

Acquisition, analysis, or interpretation of data: Helfand, Jones, D'Aquila, Tabloski, Erickson, Yue, Fong, Hshieh, Metzger, Schmitt, Boudreaux, Inouye

Drafting of the manuscript: Helfand, Jones.

Critical revision of the manuscript for important intellectual content: Helfand, D'Aquila, Yue, Fong, Tabloski, Hshieh, Metzger, Erickson, Schmitt, Boudreaux, Inouye.

Statistical analysis: Helfand.

Obtained funding: Helfand, Jones, Inouye.

Administrative, technical, or material support: Helfand, Jones, Schmitt, Inouye, Boudreaux.

Supervision: Helfand, Jones, Inouye, Boudreaux.

### **CHAPTER III – Harmonization of Four Delirium Instruments: Creating Crosswalks and the Delirium Item-Bank (DEL-IB)**

Chapter III is adapted from a manuscript in preparation for submission and included with permission not required.

#### ***Abstract***

**Objectives:** Over 30 instruments are in current, active use for delirium identification. In a recent systematic review, we recommended four commonly used and well-validated instruments for clinical and research use. The goal of this study is to harmonize the four instruments on the same metric using modern methods in psychometrics.

**Design:** Secondary data analysis from three studies, and a simulation study based on the observed data.

**Setting:** Hospitalized adults over 65 years old in the United States, Ireland, and Belgium.

**Participants:** The total sample comprised 600 participants, contributing 1,623 assessments.

**Measurements:** Confusion Assessment Method (long-form and short-form), Delirium Observation Screening Scale, Delirium Rating Scale-Revised-98 (total and severity scores), and Memorial Delirium Assessment Scale.



Results: Using item response theory, we linked scores across instruments, placing all four instruments and their separate scorings on the same metric (the propensity to delirium). Kappa statistics comparing agreement in delirium identification among the instruments ranged from 0.37-0.75, with the highest between the DRS-R-98 total score and MDAS. After linking scores, we created a harmonized item bank, called the Delirium Item Bank (DEL-IB), consisting of 50 items. The DEL-IB allowed us to create six crosswalks, which easily obtain equivalent scores across instruments.

Conclusions: Based on our results, individual instrument scores can be directly compared to aid in clinical decision-making, and quantitatively combined in meta-analyses.

## ***Introduction***

Delirium is a syndrome characterized by an acute onset of inattention, disorientation, and other cognitive disturbances that disproportionately impacts adults age 65 and older (12). It has substantial public health impact with occurrence in over 2.6 million older Americans, accounting for over \$164 billion in healthcare expenditures annually (16). The effects can persist long after onset, leading to prolonged hospitalization and increased risk of dementia and death (15, 16). However, in contrast to its large impact on public health, delirium remains understudied (16, 69). Although there are methods to prevent delirium (18), there remains no consensus on effective treatments (51).

One potential problem that has stymied progress in the delirium field is the fact that there are many methods for the identification of delirium with no direct approach to quantify their agreement or correspondence. Measures for identification of delirium include instruments for screening and/or diagnosing delirium. The lack of a unified approach for identification has led to over 30 instruments in active use for screening or diagnosis of delirium (8). Delirium instruments in active use offer varying assessments that question different signs and symptoms inherent to delirium. In our recent systematic review, we selected the Confusion Assessment Method (CAM), Delirium Observation Screening Scale (DOSS), Delirium Rating Scale-Revised-98 (DRS-R-98), and Memorial Delirium Assessment Scale (MDAS) as the instruments that were the most

commonly used, had high quality psychometric validity data, and allowed for rating of the reference standard Diagnostic and Statistical Manual of Mental Disorders (DSM)-5 criteria (8).

The goal of this paper is to describe the harmonization of four delirium identification instruments from our systematic review: the CAM, DOSS, DRS-R-98, and MDAS. We used three data sources, each of which administered multiple instruments to participants with overlapping instruments across data sources (i.e., common instruments across data sets), which allowed for harmonization. We used modern psychometric methods in portraying how well these instruments assess the same underlying concept and describe characteristics of the measurement of delirium identification in three samples of older hospitalized patients. These methods are used to create an item bank, which is a dataset containing each individual instrument's items and their corresponding estimated population level item response theory (IRT) parameters. This item bank is called the Delirium Item Bank (DEL-IB).

## ***Methods***

### Study Samples

We used three datasets to conduct this study. The first study is the Better Assessment of Illness (BASIL) study, which has been described previously (70).

In brief, BASIL is a prospective cohort study of English-speaking, hospitalized

adults age 70 and older living in or near Boston, MA, USA. The study enrolled 352 patients between October 20, 2015 and March 15, 2017 who underwent a total of 1,187 individual assessments (1-15 daily assessments per participant) (70). Each study participant was assessed for delirium with the following four instruments: the Confusion Assessment Method (CAM: short-form and long-form), the Memorial Delirium Assessment Scale (MDAS), the Delirium Rating Scale-Revised-98 (DRS-R-98: severity score-first 13 items), and Delirium Observation Screening Scale (DOSS: 4 items, specifically items 2, 3, 4, and 12).

The second dataset comes from Detroyer et al. (71). Patients were recruited from a palliative care unit in a university hospital in Belgium. A total of 48 patients were recruited, who underwent a total of 113 individual assessments. Each patient was examined up to three times a day during the first 10 days of their hospitalization and was assessed with the full 13-item DOSS and CAM short-form.

The third dataset comes from Adamis et al. (72). Patients over age 70 admitted to acute medical teams in a regional hospital in Ireland were recruited. A total of 200 patients were enrolled, who underwent a total of 323 individual assessments. Each patient was assessed using the DRS-R-98 (total score—all 16 items), and CAM short-form. Additionally, this study collected data on DSM-5 and DSM-IV defined delirium as their reference standard. Thus, a total of 4 instruments were

used across these 3 datasets; these instruments and items are shown in **Figure 3.1**.

**Figure 3.1. Data Structure and Models**

*Data structure*

<b>BASIL</b>	<b>Detroyer et al.</b>	<b>Adamis et al.</b>
CAM Short-form (1)	<i>CAM Short-form (1)</i>	<i>CAM Short-form (1)</i>
CAM Short-form (2)	<i>CAM Short-form (2)</i>	<i>CAM Short-form (2)</i>
CAM Short-form (3)	<i>CAM Short-form (3)</i>	<i>CAM Short-form (3)</i>
CAM Short-form (4)	<i>CAM Short-form (4)</i>	<i>CAM Short-form (4)</i>
CAM Short-form (5)	<i>CAM Short-form (5)</i>	<i>CAM Short-form (5)</i>
MDAS (1)	<i>NA</i>	<i>NA</i>
MDAS (2)	<i>NA</i>	<i>NA</i>
MDAS (3)	<i>NA</i>	<i>NA</i>
MDAS (4)	<i>NA</i>	<i>NA</i>
MDAS (5)	<i>NA</i>	<i>NA</i>
MDAS (6)	<i>NA</i>	<i>NA</i>
MDAS (7)	<i>NA</i>	<i>NA</i>
MDAS (8)	<i>NA</i>	<i>NA</i>
MDAS (9)	<i>NA</i>	<i>NA</i>
MDAS (10)	<i>NA</i>	<i>NA</i>
DRS-R-98 (1)	<i>NA</i>	<i>DRS-R-98 (1)</i>
DRS-R-98 (2)	<i>NA</i>	<i>DRS-R-98 (2)</i>
DRS-R-98 (3)	<i>NA</i>	<i>DRS-R-98 (3)</i>
DRS-R-98 (4)	<i>NA</i>	<i>DRS-R-98 (4)</i>
DRS-R-98 (5)	<i>NA</i>	<i>DRS-R-98 (5)</i>
DRS-R-98 (6)	<i>NA</i>	<i>DRS-R-98 (6)</i>
DRS-R-98 (7)	<i>NA</i>	<i>DRS-R-98 (7)</i>
DRS-R-98 (8)	<i>NA</i>	<i>DRS-R-98 (8)</i>

DRS-R-98 (9)	NA	<i>DRS-R-98 (9)</i>
DRS-R-98 (10)	NA	<i>DRS-R-98 (10)</i>
DRS-R-98 (11)	NA	<i>DRS-R-98 (11)</i>
DRS-R-98 (12)	NA	<i>DRS-R-98 (12)</i>
DRS-R-98 (13)	NA	<i>DRS-R-98 (13)</i>
NA	NA	<i>DRS-R-98 (14)</i>
NA	NA	<i>DRS-R-98 (15)</i>
NA	NA	<i>DRS-R-98 (16)</i>
NA	DOSS (1)	NA
DOSS (2)	<i>DOSS (2)</i>	NA
DOSS (3)	<i>DOSS (3)</i>	NA
DOSS (4)	<i>DOSS (4)</i>	NA
NA	DOSS (5)	NA
NA	DOSS (6)	NA
NA	DOSS (7)	NA
NA	DOSS (8)	NA
NA	DOSS (9)	NA
NA	DOSS (10)	NA
NA	DOSS (11)	NA
DOSS (12)	<i>DOSS (12)</i>	NA
NA	DOSS (13)	NA
<i>CAM Short-form (1) = CAM Long-form (1)</i>	NA	NA
<i>CAM Short-form (2) = CAM Long-form (2)</i>	NA	NA
<i>CAM Short-form (3) = CAM Long-form (3)</i>	NA	NA
<i>CAM Short-form (4) = CAM Long-form (4)</i>	NA	NA
<i>CAM Short-form (5) = CAM Long-form (5)</i>	NA	NA
CAM Long-form (6)	NA	NA
CAM Long-form (7)	NA	NA
CAM Long-form (8)	NA	NA
CAM Long-form (9)	NA	NA
CAM Long-form (10)	NA	NA

<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>	<i>Model 4</i>	<i>Model 5</i>	<i>Model 6</i>	<i>Model 7</i>
<b>BASIL</b>	<b>BASIL</b>	<b>BASIL</b>	<b>BASIL</b>	<b>Detroyer</b>	<b>Adamis</b>	<b>BASIL</b>
CAM Short-form (1)	CAM Short-form (1)	CAM Short-form (1)	CAM Short-form (1)	CAM Short-form (1)		CAM Short-form (1) = CAM Long-form (1)
CAM Short-form (2)	CAM Short-form (2)	CAM Short-form (2)	CAM Short-form (2)	CAM Short-form (2)		CAM Short-form (2) = CAM Long-form (2)
CAM Short-form (3)	CAM Short-form (3)	CAM Short-form (3)	CAM Short-form (3)	CAM Short-form (3)		CAM Short-form (3) = CAM Long-form (3)
CAM Short-form (4)	CAM Short-form (4)	CAM Short-form (4)	CAM Short-form (4)	CAM Short-form (4)		CAM Short-form (4) = CAM Long-form (4)
CAM Short-form (5)	CAM Short-form (5)	CAM Short-form (5)	CAM Short-form (5)	CAM Short-form (5)		CAM Short-form (5) = CAM Long-form (5)
	MDAS (1)	DRS-R-98 (1)	DOSS (2)	DOSS (1)	DRS-R-98 (1)	CAM Long-form (6)
	MDAS (2)	DRS-R-98 (2)	DOSS (3)	DOSS (2)	DRS-R-98 (2)	CAM Long-form (7)
	MDAS (3)	DRS-R-98 (3)	DOSS (4)	DOSS (3)	DRS-R-98 (3)	CAM Long-form (8)
	MDAS (4)	DRS-R-98 (4)	DOSS (12)	DOSS (4)	DRS-R-98 (4)	CAM Long-form (9)
	MDAS (5)	DRS-R-98 (5)		DOSS (5)	DRS-R-98 (5)	CAM Long-form (10)
	MDAS (6)	DRS-R-98 (6)		DOSS (6)	DRS-R-98 (6)	
	MDAS (7)	DRS-R-98 (7)		DOSS (7)	DRS-R-98 (7)	
	MDAS (8)	DRS-R-98 (8)		DOSS (8)	DRS-R-98 (8)	
	MDAS (9)	DRS-R-98 (9)		DOSS (9)	DRS-R-98 (9)	
	MDAS (10)	DRS-R-98 (10)		DOSS (10)	DRS-R-98 (10)	
		DRS-R-98 (11)		DOSS (11)	DRS-R-98 (11)	
		DRS-R-98 (12)		DOSS (12)	DRS-R-98 (12)	
		DRS-R-98 (13)		DOSS (13)	DRS-R-98 (13)	
					DRS-R-98 (14)	
					DRS-R-98 (15)	
					DRS-R-98 (16)	

BASIL = Better Assessment of Illness study; Yellow = CAM = Confusion Assessment Method; Red = DOSS = Delirium Observation Screening Scale; Blue = DRS-R-98 = Delirium Rating Scale-Revised-98; Green = MDAS = Memorial Delirium Assessment Scale

The darker shaded, non-italicized cells had their parameter estimates freely estimated. The lighter shaded, italicized cells had their parameter estimates held constant across different models to link the instruments together in performing harmonization.



### Four Harmonized Delirium Identification Instruments

The CAM, DRS-R-98, MDAS, and DOSS are all used to rate delirium signs and symptoms either following brief interviews or based upon observations by clinicians. While the instruments encompass similar features of delirium, they each have unique characteristics.

The CAM long-form consists of 10 items based on the DSM-III-R criteria for delirium. The CAM is the only one of these instruments that can be scored using a diagnostic algorithm, rather than an additive score. The algorithm requires the presence of acute onset and/or fluctuation, inattention, and either disorganized thinking or altered level of consciousness (1). The items of the CAM algorithm are operationalized to make the CAM short-form. The long-form additionally includes the following items: disorientation, memory impairment, perceptual disturbances, psychomotor agitation, psychomotor retardation, and altered sleep-wake cycle. Most CAM features are scored on a three-point scale to give a total score on the CAM. Each item is rated 0 (absent), 1 (mild), or 2 (marked), except acute onset or fluctuation, which are rated 0 (absent) or 1 (present). For our analysis, we used the scoring from the worksheets for the CAM long (scored 0-20) and short (scored 0-5) forms (73, 74). In the CAM long-form, we coded acute onset and fluctuating course as separate variables; thus, our scoring ranges from 0-20. However, Inouye and colleagues have also described the CAM-S long-form

scoring using a single item for acute onset and fluctuating course and the score range was 0-19 (4, 75).

The DRS-R-98 instructs assessors to use any accessible information source including chart review, nurses, and family to rate and identify delirium according to 13 items that characterize severity and an additional 3 diagnostic items. It can then be used to score just the severity items, or both the severity items and diagnostic items combined. The severity items are in order: sleep/wake cycle disturbance, perceptual disturbances and hallucinations, delusions, lability of affect, language, thought process abnormalities, motor agitation, motor retardation, orientation, attention, short-term memory, long-term memory, and visuospatial ability (76). The diagnostic items are temporal onset of symptoms, fluctuation of symptom severity, and physical disorder. The ratings for each item range from 0 (no impairment) to 3 (severe impairment), except for fluctuation of symptom severity and physical disorder, which are both rated 0 to 2. The DRS-R-98 total score (16 items) ranges from 0 to 46 with an author-defined cut score of 17.75 for defining presence of delirium and the DRS-R-98 severity score (13 items) ranges from 0 to 39 with an author-defined cut score of 15.25 for defining presence of delirium.

The DOSS instructs assessors to rate delirium using 13 items on a binary scale with scores ranging from 0 to 13 (38, 77, 78). Scores  $\geq 3$  indicate the patient

likely has delirium (38, 77, 78). The instrument was designed to be administered by bedside nurses once per shift through observation of verbal and nonverbal behaviors. Items include: dozes off during conversation or activities; is easily distracted by stimuli from the environment; maintains attention to conversation or action; does not finish question or answer; gives answers that do not fit the question; reacts slowly to instructions; thinks they are somewhere else; knows which part of the day it is; remembers recent events; is picking, disorderly, restless, pulls intravenous (IV) tubing, feeding tubes, catheters etc.; is easily or suddenly emotional; and sees/hears things which are not there (38).

The MDAS uses a four-point scale (0 to 3) for each of its 10 items (24). The instrument items were selected based on DSM-IV criteria and include in order: reduced level of consciousness, disorientation, impaired short-term memory, impaired digit span, reduced ability to maintain and shift attention, disorganized thinking, perceptual disturbance, delusions, psychomotor activity, and sleep-wake cycle disturbances. The MDAS ranges from 0 to 30 with an author-defined cut score of 13 for defining presence of delirium.

#### Data Analysis: Harmonization

Harmonization is a form of test score linking that enables the transformation of data from multiple sources in a comparable way such that they can be treated as equivalent (45, 47). Harmonization is a technique that has been used to link

datasets including the Health and Retirement Study to other similar cross-national studies (79). Our approach involved the use of item response theory (IRT)-based co-calibration of the four instruments. Instrument metrics are linked through the presence of common (linking) items available across studies. Linking items are items that are or can be assumed to be equivalent across studies. Our approach involves the assumption that all instruments measure the same underlying trait.

IRT describes a large body of latent variable models used to describe relationships between the latent trait that underlies the instrument and the responses to the individual items that comprise the instrument (i.e., the item responses). In our analysis, the latent trait is conceptualized as the propensity to delirium; the item responses are generated from the individual questions on the delirium identification instruments, which assess the signs and symptoms of delirium. We use IRT to harmonize all instruments from the different datasets, allowing for their comparison on the same metric. We fit a graded response model, which estimates a discrimination parameter and boundary (difficulty) parameters between response categories. The discrimination parameter describes how well each item separates individuals of low and high levels of the latent trait (27, 43). The boundary, or difficulty, parameters identify the level on the latent trait at which individuals are more likely to be in the next higher

response category (27, 43, 80). The collection of item parameters for all items comprises the item bank.

To perform the statistical harmonization, we used IRT-based generalized structural equation models, and chose unidimensional factor models since their fit was considered adequate and appropriate for our aims. Then, we matched instruments on the same metric using a combination of the anchor-test design and common-person design (43). In the anchor-test design, common items are administered to different study populations. For example, in our study, four questions of the DOSS were given in the BASIL sample, while all of the questions were given in the Detroyer et al. sample, allowing us to link the instruments. In the common-person design, common instruments are given to different study populations. In our study, the CAM short-form was given across every dataset. In total, we fit seven different models using the generalized structural equation modeling procedures in Stata (version 16.1, College Station, Texas) to estimate item parameters. The structure of our models is shown in **Figure 3.1**. To accomplish the harmonization, we constrained (i.e., held constant) item parameters on items that were in common across different models to link all the instruments.

To summarize our approach, we have shown each of the designs of our models in the steps to complete the full harmonization of each instrument in **Figure 3.1**.

We began with the CAM short-form items being held constant in the models, since these were constant across all datasets. We first estimated an IRT model to find the item parameters in the CAM short-form using participants from the BASIL study. Second, we held the CAM short-form parameters from the first model constant, and freely estimated all MDAS items from the BASIL study. Third, we again held the CAM short-form parameters from the first model constant, and freely estimated the 13 items in the DRS-R-98 severity score from the BASIL study. Fourth, we held the CAM short-form parameters constant, and freely estimated the 4 items from the DOSS found in the BASIL study (specifically DOSS items 2, 3, 4, and 12). Fifth, we held the parameters from CAM short-form and 4 items from the DOSS found in the BASIL study constant, and freely estimated the remaining DOSS items from the Detroyer et al. dataset. Sixth, we held the parameters from CAM short-form and 13 items in the DRS-R-98 severity score found in the BASIL study constant, and freely estimated the remaining 3 items from the DRS-R-98 total score found in the Adamis et al. dataset. Seventh, we held the CAM short-form parameters constant, and freely estimated the CAM long-form items from the BASIL study.

In all datasets, items that were coded as “uncertain” or “don’t know” were set to missing. If at least one item was non-missing for a person-visit, that person-visit was included in the models.

### Simulation Methods

To generate our crosswalks between instruments, we used simulation procedures based on our observed data. For this simulation, our goal was to generate a single large sample of persons and their item responses to all of the delirium assessment items in our item bank, the DEL-IB. We wanted to generate a hypothetical cohort that was large enough (>100 times the size of our combined cohort in this study) to have demonstrated scores on all instrument items included in the DEL-IB. Boundary parameter estimates not observed in the real data were extrapolated from observed parameter estimates. We created the simulated dataset using the R-based program Firestar (81). We input our item parameters already found in the DEL-IB and had the program create a simulated sample size of  $N=100,001$ , with each participant responding to each item in the DEL-IB. The underlying latent trait was weighted to a normal distribution with an assumed mean of 0 and standard deviation of 1. We used these responses to generate expected score characteristic curves and crosswalks for all measures. Expected score characteristic curves are the curves made from the parameter estimates in the DEL-IB.

Crosswalks are a representation of equivalent scores on different instruments. We used similar methods to create reliability or measurement precision curves, which reveal the level of accuracy with which a given instrument measures the underlying latent trait.

## Results

The total sample size was 600 participants, who contributed 1623 unique assessments. **Table 3.1** describes the study characteristics across the three studies. The BASIL study and Adamis et al. study had study populations with mean age over 80 years and balanced participant genders. The Detroyer et al. study had a younger study population with median age of 72, and 38% of the study population was female. The rates of delirium across the studies based on CAM criteria, which was in common across all the studies, ranged from 17%-25%. Notably, the Adamis et al. study had a high prevalence of patients with dementia (63%).

**Table 3.1. Baseline characteristics of the three datasets**

	BASIL (N=352)	Detroyer et al. (N=48)	Adamis et al. (N=200)
Age, years, mean (SD) or median (IQR)	80.3 (6.8)	72 (67.25; 78)	81.1 (6.5)
Female sex, n (%)	203 (58)	18 (38)	100 (50)
Non-white race, n (%)	48 (14)	NR	NR
Years of education, mean (SD)	14.5 (3.0)	NR	NR
Married, n (%)	139 (40)	26 (54)	NR
Lives alone, n (%)	135 (39)	7 (15)	NR
Lives in nursing home, n (%)	13 (3.7)	1 (2.1)	NR
Dementia/previous history of cognitive impairment, n (%)	101 (29)	NR	126 (63)
CAM delirium (ever), n (%)	88 (25)	11 (23)	34 (17)

NR = not reported; SD = standard deviation; IQR = interquartile range; BASIL = Better Assessment of Illness Study; CAM = confusion assessment method



**Table 3.2** are the kappa statistics of agreement in delirium identification between the instruments using their author-described definitions. The range in kappa statistics was 0.37-0.75. This range describes agreement that is considered fair to substantial (82). The highest levels of agreement were between the DRS-R-98 total score and MDAS with kappa=0.75.

**Table 3.2. Kappa statistics of delirium identification between CAM (short), DOSS, DRS-R-98, MDAS**

	CAM (short)	DOSS	MDAS
DOSS	.61	---	---
MDAS	.56	.37	---
DRS-R-98 (severity)	.70	.53	.69
DRS-R-98 (total)	.63	.44	.75

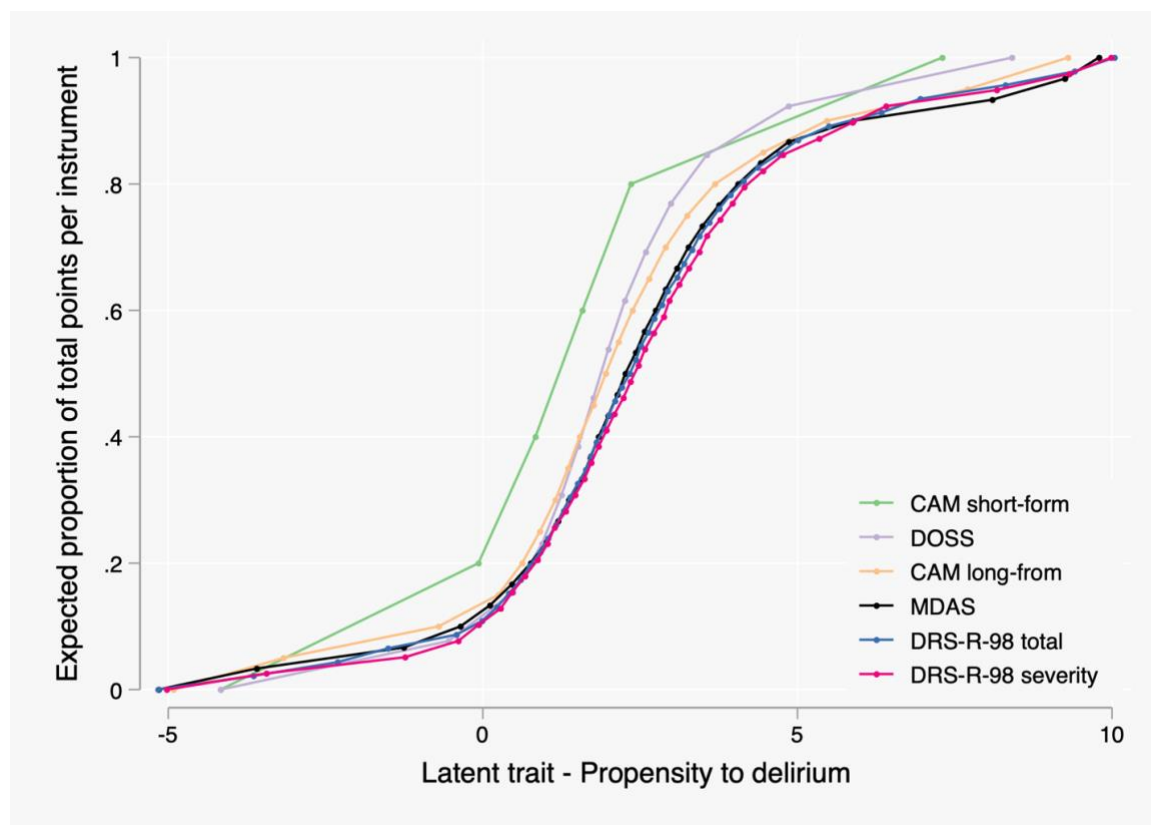
CAM = Confusion Assessment Method (short-form); DOSS = Delirium Observation Screening Scale; DRS-R-98 = Delirium Rating Scale-Revised-98; MDAS = Memorial Delirium Assessment Scale

Kappa values can be interpreted as slight for .01-.20, fair for .21-.40, moderate for .41-.60, substantial for .61-.80, and almost perfect agreement for .81-1.0 (83).

In **Figure 3.2**, we show the expected score characteristic curves for each of the instruments. In **Figure 3.2**, we display the expected score on each instrument at different levels along the latent trait. At low levels of the latent trait, there are few endorsed signs and symptoms of delirium, leading to low scores. As the latent

trait increases and participants have a higher propensity to delirium, more signs and symptoms are endorsed and scores increase.

**Figure 3.2. Expected score characteristic curves of each delirium identification instrument**

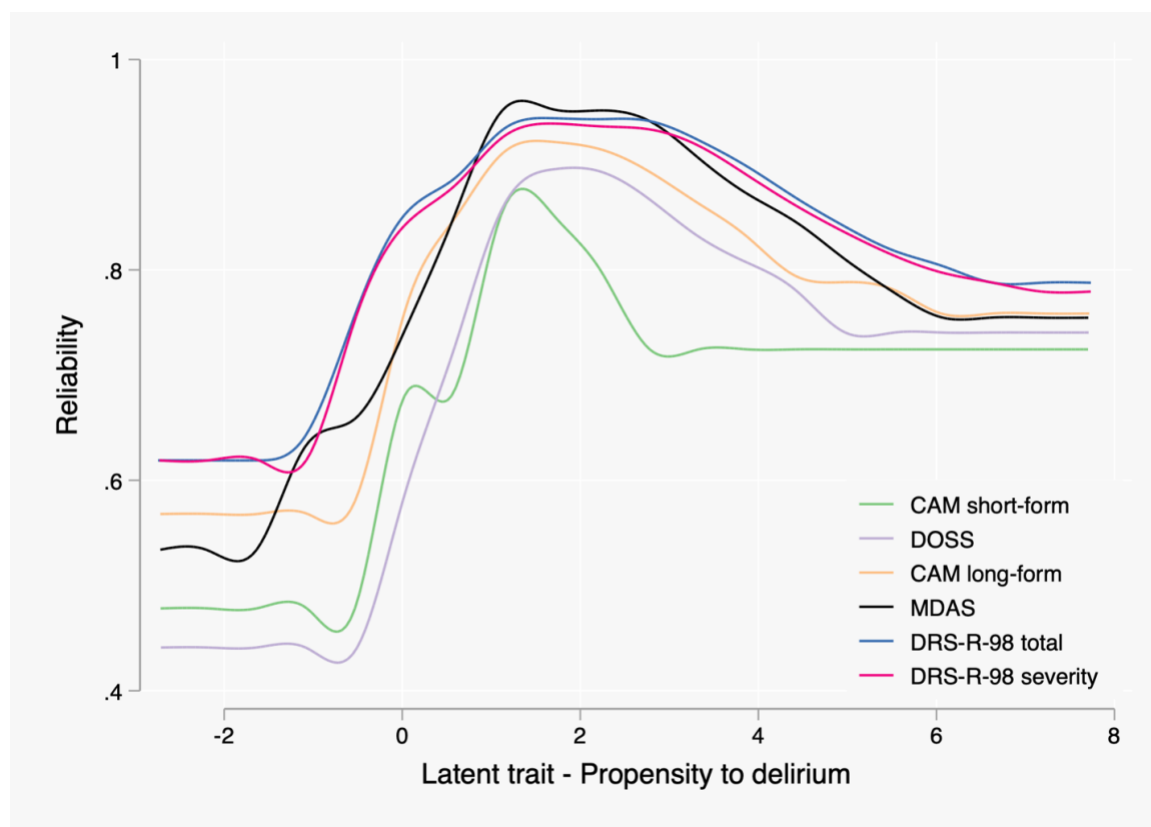


CAM = Confusion Assessment Method; DOSS = Delirium Observation Screening Scale; DRS-R-98 = Delirium Rating Scale-Revised-98; MDAS = Memorial Delirium Assessment Scale

Figure Legend: Expected score characteristic curves of each delirium identification instrument are shown. Each curve shows the expected proportion of total points a participant would have on each instrument across the latent trait, propensity to delirium.

**Figure 3.3** shows the reliability of each delirium identification instrument. These curves show the varying reliability of each instrument across different levels of the latent trait, propensity to delirium. Each of the curves has a peak reliability that falls on the latent trait at roughly the same level, between 1.5-2.0.

**Figure 3.3. Reliability of each delirium identification instrument**



CAM = Confusion Assessment Method; DOSS = Delirium Observation Screening Scale; DRS-R-98 = Delirium Rating Scale-Revised-98; MDAS = Memorial Delirium Assessment Scale

Figure Legend: Measurement reliability or precision of each different delirium identification instrument is displayed across the latent trait, propensity to delirium.

As shown in **Figure 3.2**, by aligning the expected score characteristic curves of each instrument on the same latent trait, we are able to make equivalent scores across each instrument to generate the crosswalks, as shown in **Table 3.3 through Table 3.8**. For integer scores, crosswalks only work in a single direction; thus, it is important to use the proper one when comparing or transforming scores from one instrument to another. Each table can be read by starting with the source instrument in the first column and moving along the row to see the equivalent score on each of the other 5 instruments, as well as where the participant would fall on the latent trait, propensity to delirium.

**Table 3.3. DOSS crosswalk**

Source Instrument	Equivalent Scores					
	DOSS	CAM Short-form	CAM Long-form	MDAS	DRS-R-98 Total	DRS-R-98 Severity
0	0	0	0	1	0	-4.3
1	1	2	3	4	3	-0.3
2	1	3	5	7	5	0.6
3	2	5	6	10	8	1.0
4	3	7	8	13	11	1.2
5	3	8	10	15	12	1.5
6	3	9	11	17	14	1.8
7	4	10	13	20	16	2.0
8	4	12	15	22	18	2.2
9	4	13	17	25	21	2.6
10	5	14	20	29	24	3.0
11	5	16	22	33	28	3.5
12	5	18	26	40	33	4.7
13	5	19	29	44	38	8.5

CAM = Confusion Assessment Method; DOSS = Delirium Observation Screening Scale; DRS-R-98 = Delirium Rating Scale-Revised-98; MDAS = Memorial Delirium Assessment Scale

**Table 3.4. CAM Short-form crosswalk**

Source Instrument	Equivalent Scores					
	CAM Long-form	DOSS	MDAS	DRS-R-98 Total	DRS-R-98 Severity	Latent Trait
CAM Short-form						
0	0	0	0	1	0	-4.2
1	2	1	3	5	4	0
2	4	2	6	9	7	0.9
3	8	5	10	15	13	1.6
4	12	8	15	23	19	2.3
5	19	13	28	44	37	7.3

CAM = Confusion Assessment Method; DOSS = Delirium Observation Screening Scale; DRS-R-98 = Delirium Rating Scale-Revised-98; MDAS = Memorial Delirium Assessment Scale

**Table 3.5. CAM Long-form crosswalk**

Source Instrument	Equivalent Scores					
	CAM Short-form	DOSS	MDAS	DRS-R-98 Total	DRS-R-98 Severity	Latent Trait
CAM Long-form						
0	0	0	0	0	0	-5.2
1	0	0	1	1	1	-2.8
2	0	1	2	3	3	-0.3
3	1	1	4	6	5	0.3
4	1	2	5	8	6	0.7
5	2	3	6	10	8	0.9
6	2	3	8	12	10	1.2
7	3	4	8	13	11	1.4
8	3	5	10	15	13	1.5
9	3	6	11	17	14	1.8
10	4	7	13	20	16	2.0
11	4	8	14	22	18	2.1
12	4	8	16	23	19	2.4
13	4	10	18	26	21	2.6
14	5	10	20	29	24	2.9
15	5	11	21	31	26	3.2
16	5	11	23	35	29	3.6
17	5	12	25	38	32	4.3
18	5	13	27	41	35	5.2
19	5	13	28	44	37	7.5

20

5

13

29

45

38

9.5

---

CAM = Confusion Assessment Method; DOSS = Delirium Observation Screening Scale; DRS-R-98 = Delirium Rating Scale-Revised-98; MDAS = Memorial Delirium Assessment Scale



**Table 3.6. DRS-R-98 Severity crosswalk**

Source Instrument	Equivalent Scores					DRS-R-98 Total	Latent Trait
	DRS-R-98 Severity	CAM Short-form	CAM Long-form	DOSS	MDAS		
0	0	0	0	0	0	0	-5.3
1	0	1	0	1	1	1	-3.2
2	0	1	0	2	2	2	-0.9
3	1	2	1	3	4	4	-0.3
4	1	2	1	3	5	5	0.0
5	1	3	1	4	6	6	0.3
6	1	3	2	4	7	7	0.5
7	2	4	2	5	8	8	0.7
8	2	5	2	6	9	9	0.9
9	2	5	3	7	11	11	1.0
10	2	6	3	8	12	12	1.2
11	3	7	4	8	13	13	1.3
12	3	8	4	9	14	14	1.5
13	3	8	5	10	16	16	1.6
14	3	9	6	11	17	17	1.7
15	4	10	6	12	19	19	1.9
16	4	10	7	13	20	20	2.0
17	4	11	7	14	21	21	2.1
18	4	12	8	15	22	22	2.2

19	4	12	8	16	23	2.4
20	4	13	9	16	25	2.5
21	4	13	9	17	25	2.6
22	5	14	10	18	27	2.7
23	5	14	10	20	29	2.9
24	5	14	10	20	29	3.0
25	5	15	10	21	31	3.1
26	5	16	11	21	32	3.3
27	5	16	11	22	33	3.4
28	5	16	11	23	34	3.5
29	5	17	12	23	36	3.7
30	5	17	12	24	37	4.0
31	5	17	12	24	37	4.1
32	5	18	12	25	39	4.4
33	5	18	12	26	40	4.7
34	5	19	13	27	41	5.2
35	5	19	13	28	42	5.7
36	5	19	13	28	43	6.1
37	5	19	13	28	44	8.0
38	5	20	13	29	45	9.5
39	5	20	13	29	45	10.3

---

CAM = Confusion Assessment Method; DOSS = Delirium Observation Screening Scale; DRS-R-98 = Delirium Rating Scale-Revised-98; MDAS = Memorial Delirium Assessment Scale

**Table 3.7. DRS-R-98 Total crosswalk**

Source Instrument	Equivalent Scores					DRS-R-98 Severity	Latent Trait
	DRS-R-98 Total	CAM Short-form	CAM Long-form	DOSS	MDAS		
0	0	0	0	0	0	0	-5.4
1	0	1	0	1	1	1	-3.4
2	0	1	0	1	1	1	-1.7
3	0	1	0	1	1	1	-1.1
4	1	2	1	3	3	3	-0.3
5	1	2	1	3	4	4	0.1
6	1	3	1	4	5	5	0.3
7	1	3	1	4	6	6	0.5
8	1	4	2	5	6	6	0.6
9	2	4	2	5	7	7	0.8
10	2	5	3	6	8	8	0.9
11	2	5	3	7	9	9	1.1
12	2	6	3	8	10	10	1.2
13	3	7	4	8	11	11	1.3
14	3	7	4	9	12	12	1.4
15	3	8	5	10	13	13	1.5
16	3	8	6	11	14	14	1.7
17	3	9	6	11	14	14	1.7
18	4	9	6	12	14	14	1.8

19	4	10	7	13	16	1.9
20	4	10	7	13	16	2.0
21	4	11	7	14	17	2.1
22	4	11	8	15	18	2.2
23	4	12	8	16	19	2.3
24	4	13	9	16	20	2.4
25	4	13	9	17	20	2.5
26	5	13	10	18	22	2.6
27	5	14	10	18	22	2.7
28	5	14	10	19	23	2.8
29	5	14	10	20	24	2.9
30	5	15	10	20	25	3.1
31	5	15	11	21	26	3.2
32	5	15	11	21	26	3.3
33	5	16	11	22	27	3.4
34	5	16	11	23	28	3.6
35	5	16	11	23	29	3.7
36	5	17	12	24	30	3.9
37	5	17	12	24	31	4.1
38	5	17	12	25	32	4.3
39	5	18	12	26	33	4.6
40	5	18	13	26	34	4.9
41	5	18	13	27	35	5.3
42	5	19	13	28	36	6.1
43	5	19	13	28	37	6.6

44	5	19	13	28	37	8.2
45	5	20	13	29	38	9.6
46	5	20	13	29	38	10.4

---

CAM = Confusion Assessment Method; DOSS = Delirium Observation Screening Scale; DRS-R-98 = Delirium Rating Scale-Revised-98; MDAS = Memorial Delirium Assessment Scale

**Table 3.8. MDAS crosswalk**

Source Instrument	Equivalent Scores						
	MDAS	CAM Short-form	CAM Long-form	DOSS	DRS-R-98 Total	DRS-R-98 Severity	Latent Trait
0	0	0	0	0	0	0	-5.4
1	0	1	1	0	1	0	-3.4
2	0	1	1	0	2	1	-1.0
3	1	2	2	1	4	3	-0.2
4	1	2	2	1	5	4	0.2
5	1	3	3	2	7	5	0.5
6	2	4	4	2	9	7	0.8
7	2	5	5	3	10	8	1.0
8	3	6	6	3	12	10	1.2
9	3	7	7	4	14	12	1.4
10	3	8	8	5	15	13	1.6
11	3	9	9	6	17	14	1.7
12	4	10	10	6	19	15	1.8
13	4	10	10	7	20	16	2.0
14	4	11	11	7	21	17	2.1
15	4	12	12	8	22	18	2.3
16	4	12	12	9	24	20	2.4
17	4	13	13	9	25	21	2.6
18	5	14	14	10	27	22	2.7

19	5	14	10	28	23	2.9
20	5	15	10	31	25	3.0
21	5	15	11	31	26	3.2
22	5	16	11	33	27	3.4
23	5	16	11	35	29	3.7
24	5	17	12	37	31	4.0
25	5	18	12	39	32	4.4
26	5	18	12	40	34	4.8
27	5	19	13	42	36	5.6
28	5	19	13	44	37	7.9
29	5	20	13	45	38	9.4
30	5	20	13	45	38	10.1

---

CAM = Confusion Assessment Method; DOSS = Delirium Observation Screening Scale; DRS-R-98 = Delirium Rating Scale-Revised-98; MDAS = Memorial Delirium Assessment Scale

## ***Discussion***

We used modern psychometric methods including IRT to harmonize the CAM, DRS-R-98, DOSS, and MDAS on the same metric. Using three independent data sets, we were able to cross-link four instruments for delirium identification, using the common CAM short-form items as an anchor. We created crosswalks of scores, putting all the instruments on the same metric using IRT approaches. Importantly, we generated the DEL-IB with 50 items, which includes individual items scores and their population-based parameter estimates. The DEL-IB will provide an important resource for future work.

Harmonization of four commonly used and well-validated instruments represents a substantial advance for the field. Currently, when studies use different delirium instruments, delirium rates may vary across studies, resulting in the potential for flawed or misleading conclusions. The DEL-IB allowed for the creation of crosswalks that permit direct comparison of the delirium identification instrument scores across the instruments we harmonized. The crosswalks will allow comparison of scores on different instruments in real time. For example, a nurse presents a patient's DOSS score to the consulting psychiatrist, who will be able to determine an equivalent score on the DRS-R-98, with which the psychiatrist may be more familiar.



Our study, by applying advanced measurement methods to compare instruments, is relatively novel within the field of delirium research. The only other known use of harmonization of delirium instruments was performed previously using only the BASIL study to harmonize the measurement of delirium severity (84). In the previous harmonization work, the BASIL study was used to harmonize the CAM (short-form and long-form), DRS-R-98 (severity score only), and MDAS (84). Thus, our study extends this work to delirium identification instruments, and now includes the CAM short and long-forms, the DRS-R-98 severity and total score, and the DOSS, using three separate datasets from different geographic regions.

There are several strengths to this study. This study used a novel approach within delirium research, namely the application of advanced psychometric methods to the three independent datasets, each examining multiple delirium identification instruments. Additionally, these datasets examine patients from the United States, Ireland, and Belgium, enhancing the generalizability of the results. Each site provided multiple ratings on a robust number of participants. The fact that each of these institutions used multiple and overlapping delirium identification measures facilitated the work. The inclusion of DSM-5 reference standard ratings helped us to quantify the propensity to delirium for our study.

There are several limitations that deserve comment. First, since each of the datasets were derived from hospitalized patients, the results might not be generalizable to non-hospital settings. Second, the data was collected using various approaches, including clinical bedside observations by nurses and clinicians at two sites, and trained lay interviewers at another sites. Both of these approaches may have varied in comparison with reference standard-quality ratings by expert clinicians, such as geriatricians or geriatric psychiatrists. Third, our comparisons are based on simulation data, instead of real data on all four instruments simultaneously administered to each patient within a single study that might yield different or stronger psychometric evidence.

Crosswalks will allow comparison of equivalent delirium rates across different studies and enable pooling of data from multiple studies, regardless of the delirium identification measure used. Such pooling will facilitate combining of data across multiple studies for meta-analyses and creation of big data resources with integrative analyses of pooled data to advance studies in omics, delirium pathophysiology, machine learning or other areas requiring large samples. Future directions include delving into applications of the created DEL-IB, such as comparing author-defined cut-points for case identification. Additionally, the DEL-IB could be used to create new instruments to advance the field.

## ***Acknowledgment***

### Published Manuscript Co-Author Contributions

Mr. Helfand conceived of the project, collected the search, did all analyses, wrote the manuscript, created all tables and figures. Mr. Helfand and Dr. Jones had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Other contributions include:

Concept and design: Helfand, Jones, Inouye, Boudreaux

Acquisition, analysis, or interpretation of data: Helfand, Jones, Boudreaux, Inouye, Detroyer, Milisen, Adamis.

Drafting of the manuscript: Helfand, Jones.

Critical revision of the manuscript for important intellectual content: Helfand, Metzger, Detroyer, Milisen, Adamis, Boudreaux, Inouye.

Statistical analysis: Helfand, Jones.

Obtained funding: Helfand, Jones, Inouye.

Administrative, technical, or material support: Helfand, Jones, Inouye, Boudreaux.

Supervision: Helfand, Jones, Inouye, Boudreaux.

## CHAPTER IV – Delirium Item Bank (DEL-IB): Utilization to Evaluate and Create Delirium Instruments

Chapter IV is adapted from a manuscript in preparation for submission and included with permission not required.

### ***Abstract***

**Background:** The large number of heterogeneous instruments in active use for identification of delirium prevents direct comparison of studies and the ability to combine results. In a recent systematic review we performed, we recommended four commonly used and well-validated instruments and subsequently harmonized them on the same metric using advanced psychometric methods to develop an item bank, the Delirium Item Bank (DEL-IB).

**Objectives:** The goal of the present study is to find optimal cut-points on each instrument and to demonstrate use of the DEL-IB to create new instruments.

**Methods:** We used a secondary analysis and simulation study based on data from three previous studies of hospitalized older adults (age 65+ years) in the United States, Ireland, and Belgium. The combined dataset included 600 participants, contributing 1,623 delirium assessments. The measurements included the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) diagnostic criteria for delirium, Confusion Assessment Method (CAM)

(long-form and short-form), Delirium Observation Screening Scale (DOSS), Delirium Rating Scale-Revised-98 (DRS-R-98) (total and severity scores), and Memorial Delirium Assessment Scale (MDAS).

**Results:** We identified different cut-points for each instrument to optimize sensitivity or specificity, and Youden's J statistic, and compared instrument performance at each cut-point to the author-defined cut-point. For example, the cut-point on the MDAS at Youden's J statistic was at a sum score of 6 with 89% sensitivity and 79% specificity. Then, using the DEL-IB, we created four example instruments (two short forms and two long forms) and evaluated their performance characteristics. In the first example short form instrument, the cut-point at Youden's J statistic was at a sum score of 3 with 90% sensitivity, 81% specificity, 30% positive predictive value (PPV), and 99% negative predictive value (NPV).

**Conclusion:** We used the DEL-IB to better understand the psychometric performance of 6 current delirium identification instruments and scorings, and demonstrated its use to create new instruments. Ultimately, we hope the DEL-IB might be used to create optimized delirium identification instruments and to spur the development of a unified approach to identify delirium.

## ***Introduction***

Delirium is a public health problem that disproportionately impacts older adults. Delirium is estimated to occur in over 2.6 million older (age 65+ years) Americans annually, and accounts for over \$164 billion in healthcare expenditures (16). Unfortunately, despite its large public health impact, delirium remains understudied (16, 69). Clinically, delirium is characterized by an acute onset of inattention, disorientation, and other cognitive disturbances and is diagnosed based on clinical observations. Its effects can persist beyond the acute event leading to prolonged hospitalization, producing an increased risk of dementia and death (15, 16). Fortunately, effective approaches have been developed to prevent delirium (18). However, due to the reliance on bedside clinical diagnosis without specific laboratory markers or radiographic evidence, there is no consensus on a single, effective approach for delirium identification (51).

This lack of consensus has led to the use of a large number of instruments for identification of delirium, which in turn, has hampered progress of the field. Many of these instruments have been created without full understanding about their performance characteristics across different populations, or of their agreement with each other. There are at least 30 instruments in current use for identifying delirium (e.g., for screening or diagnosis purposes) and each of these instruments provide varying degrees of coverage of delirium domains (8). Based on procedures outlined in our recent systematic review [See Chapter II], we

selected the Confusion Assessment Method (CAM), Delirium Observation Screening Scale (DOSS), Delirium Rating Scale-Revised-98 (DRS-R-98), and Memorial Delirium Assessment Scale (MDAS) as the instruments that were the most commonly used, that had high quality psychometric validity data, and that best fulfilled Diagnostic and Statistical Manual of Mental Disorders (DSM)-5 criteria for delirium (8).

Following the systematic review, we harmonized the four selected instruments on the same metric using modern methods in psychometrics to develop a harmonized item bank, the Delirium Item Bank (DEL-IB), and to create crosswalks between the scores on all four instruments [See Chapter III]. We used three separate datasets (70-72), each containing multiple instruments administered to participants, which overlapped and allowed for harmonization of the items on the same metric, that is, the propensity to delirium. An item bank is a dataset that contains each item on each instrument, along with their estimated population level item response theory (IRT) parameters. Crosswalks provide an easy-to-use guide with corresponding scores on different instruments and can be readily used to cross-reference scores in real time across multiple instruments.

The goals of the present manuscript are twofold. First, we wanted to determine the cut-points that would best identify delirium in comparison with a common reference standard across all instruments. Second, we wanted to use the DEL-IB

to create new instruments and to demonstrate their performance characteristics using the selected cut-points. Thus, we aimed to demonstrate how the use of the DEL-IB can be used to develop and evaluate multiple new delirium instruments to advance the field.

## ***Methods***

### Study Samples.

We previously described the study samples and the preliminary creation of the DEL-IB [See Chapter III]. Briefly, we used data from three studies: Adamis et al. (n=200) (72), BASIL (Better Assessment of Illness Study) (n=352) (70), and Detroyer et al. (n=48) (71) each administering multiple and at least partially overlapping delirium identification instruments to hospitalized adults age 65 years and older. The total sample for the present analysis included 600 participants, contributing 1,623 delirium assessments. The instruments included across the studies are: Confusion Assessment Method (CAM) (long-form and short-form), Delirium Observation Screening Scale (DOSS), Delirium Rating Scale-Revised-98 (DRS-R-98) (total and severity scores), and Memorial Delirium Assessment Scale (MDAS).

### Overall Analytic Approach.

We used item response theory (IRT) to perform statistical harmonization, to select cut-points, and to guide creation of new instruments. Statistical



harmonization provides a quantitative approach to cross-link each item of each instrument on the same latent trait metric, in this case, propensity to delirium. Taken together, the items and their parameter estimates comprised the DEL-IB, created in our prior study [See Chapter III], which serves as the foundation for the present study.

Our first step in the present analysis was to identify cut-points on the four selected instruments. We started with the Adamis et al. study where the CAM and DRS-R-98 scores were related to DSM-5-defined delirium diagnoses, which was used as the reference standard (11). We then repeated these procedures using summary scores derived from the MDAS and DOSS, plus alternative versions of the CAM and DRS-R-98, by linking common items across studies and relating their performance to DSM-5-defined delirium diagnoses. We used simulation methods based on the Adamis et al. results and IRT results from our prior harmonization work [See Chapter III]. Our first goal was to determine cut-points that best identified presence (versus absence) of delirium through simulation studies on our secondary data sources. We estimated three different cut-points on each instrument: one cut-point to optimize sensitivity (>90% sensitivity), one to optimize specificity (>90% specificity), and one that balanced sensitivity and specificity at Youden's J statistic (85). We compared instrument performance at these cut-points with the author-defined cut-points and with performance on the latent trait of propensity to delirium.

To illustrate how the DEL-IB could be used to create new instruments, we generated four different examples. We aimed to first create short forms, selecting 5 items as a maximum for streamlined use in clinical practice. The first short form selected items to optimize content validity; the second short form selected items with maximum information at the optimal level on the latent trait for identifying DSM-5 delirium. Similarly, two long-forms were created with 10 items each. Again, the first long form selected items to optimize content validity; the second long form selected items with maximum information by IRT.

#### Data Analysis: Cut-points

Adamis et al. used DSM-5 criteria to diagnose delirium, and assessed each patient simultaneously using the CAM and full DRS-R-98. Since the DEL-IB contained the CAM and DRS-R-98 and Adamis et al. included these instruments alongside the reference standard diagnosis of DSM-5, we were able to generate a latent trait estimate for delirium symptom data. Using logistic regression, we developed a prediction model for DSM-5 reference standard delirium diagnosis given the latent trait estimate. We simulated a dataset of 100,001 observations applying the R-based program Firestar (81), using the existing parameter estimates across all six different instrument scorings in the DEL-IB. Then, we added the DSM-5-defined delirium diagnoses to this dataset applying the prediction model based on the Adamis et al. dataset. This allowed us to generate

scores across all six different instrument scorings in the new simulated data. Then, we related the total scores, author's cut-points, and identified cut-points that optimized sensitivity nearest to 90%, specificity nearest to 90%, and at Youden's J cut-score on each of the original instruments. We also looked at the latent trait estimate used in the item generating models in terms of sensitivity nearest to 90%, specificity nearest to 90%, and Youden's J statistic. Youden's J statistic is based on the formula:

$$J = \text{sensitivity} + \text{specificity} - 1,$$

and therefore, defined for all points along the ROC curve; the cut-point that returns the maximum J statistic is the one that maximizes both sensitivity and specificity at the same time (85). For all analyses, we used direct standardization to the BASIL sample CAM short form distribution to account for sample heterogeneity (86). We used Stata (version 16.1, College Station, Texas) in all of our analyses to develop our IRT models and receiver operator characteristic (ROC) curves.

Each instrument has different methods the author used to describe likelihood of delirium identification. The CAM (both short and long forms) defines delirium using a diagnostic algorithm (1). The DOSS defines delirium as a score  $\geq 3$  (38, 77, 78). The MDAS defines delirium as a score  $\geq 15$  (24). The DRS-R-98 severity and total defines delirium as a score  $>15.25$  and  $>17.75$  on its 13-item severity score and 16-item total score, respectively (76).

### Data Analysis: New Instruments

To demonstrate its application, we used the DEL-IB to create four new instruments, two short forms and two long forms. In creating our instruments, we wanted to select items that matched domains relevant to DSM criteria. The delirium identification domains defined from the DSM-5 diagnostic criteria, based on a previous expert panel process by our group included: acute onset, fluctuating course, inattention, disorientation, and cognitive impairment (8). The expert panel also rated other delirium identification domains covered by DSM-diagnostic criteria from earlier versions of the DSM, including DSM-III (when delirium was first codified), DSM-III-R, DSM-IV, and DSM-IV-TR. In addition to the five domains already defined, there were five additional domains identified, which included: (i) level of consciousness, (ii) disorganized thinking, (iii) psychomotor agitation, (iv) psychomotor retardation, and (v) hallucinations, perceptual disorder or distortion (8). Based on the previous expert panel process (8), each item of the CAM, DOSS, DRS-R-98, and MDAS items were matched to these domains.

The first example instrument is a short form (5 items) with highest content validity. To achieve this, items were selected based on the following criteria: within each of the five domains of the DSM-5-defined delirium diagnostic criteria,

we identified the item with the highest information content at a latent trait level that maximized the Youden's J statistic for DSM-5 delirium.

The statistical notion of information is defined as the inverse of precision (87). In the IRT context, item information refers to the inverse of the precision with which a particular item provides for estimating an individual's level on the underlying trait upon which the item responses are believed to be based (88). Information is operationalized as the inverse variance of an item response function and is computed on the basis of the estimated item response parameters (discrimination, difficulty or location) (89). Precision is not constant across the range of the latent trait; it is peaked at the level of the underlying trait where the difficulty or boundary parameters are located. As inverse variance estimates, information functions are additive across all items in an instrument. The sum of a set of items' information functions, known as test information, conveys the accuracy with which a set of items measures an underlying trait. Among the items in the DEL-IB, we used item information functions to identify, among the items in the DEL-IB, those that provide the most information in the region of the underlying trait that corresponds to the cut point that optimizes sensitivity and specificity for DSM-5 delirium. We also used test information functions to assess the quality of measurement of an instrument. The IRT notion of reliability can be expressed as a function of test information: reliability is the complement of the inverse of information (90). If we believe a good test for individual level decision

making would have a reliability of at least 0.90 (91), then the target test information level – particularly in regions of the latent trait important for making individual level decisions – should be at least 10.

The second example instrument is a short form that includes the five items with the highest information only, without regard for content balancing. The third example instrument is a long form (10 items) that includes one item from each of the 10 domains of delirium identification across all versions of the DSM, which was also selected by the same criteria as the first example instrument. Within each of the 10 domains of the DSM-defined delirium diagnostic criteria, we identified the item that has the highest information content at a latent trait level that maximized the Youden's J statistic with respect to DSM-5 delirium. The fourth example instrument is another long form that includes the 10 items with the highest information only, without regard for content balancing.

## ***Results***

Across all three studies there were 1623 unique assessments provided by 600 participants. The description of the study characteristics across each of the three studies is shown in **Table 4.1**. The study samples of the Adamis et al. and BASIL study each had comparable rates on participant sex and mean age over 80 years, while the Detroyer et al. study, with a smaller sample size (n=48), had 38% women with a median age of 72 years. The Adamis et al. sample had a high

prevalence (63%) of patients with dementia. The prevalence of CAM-defined delirium across the studies ranged from 17%-25%. The Adamis et al. sample, which provided the basis for the simulation study, had a 13% prevalence of delirium by DSM-5 criteria. This lower prevalence of delirium was adjusted by use of direct standardization techniques, as described in the methods section (86).

**Table 4.1. Baseline characteristics of the three datasets**

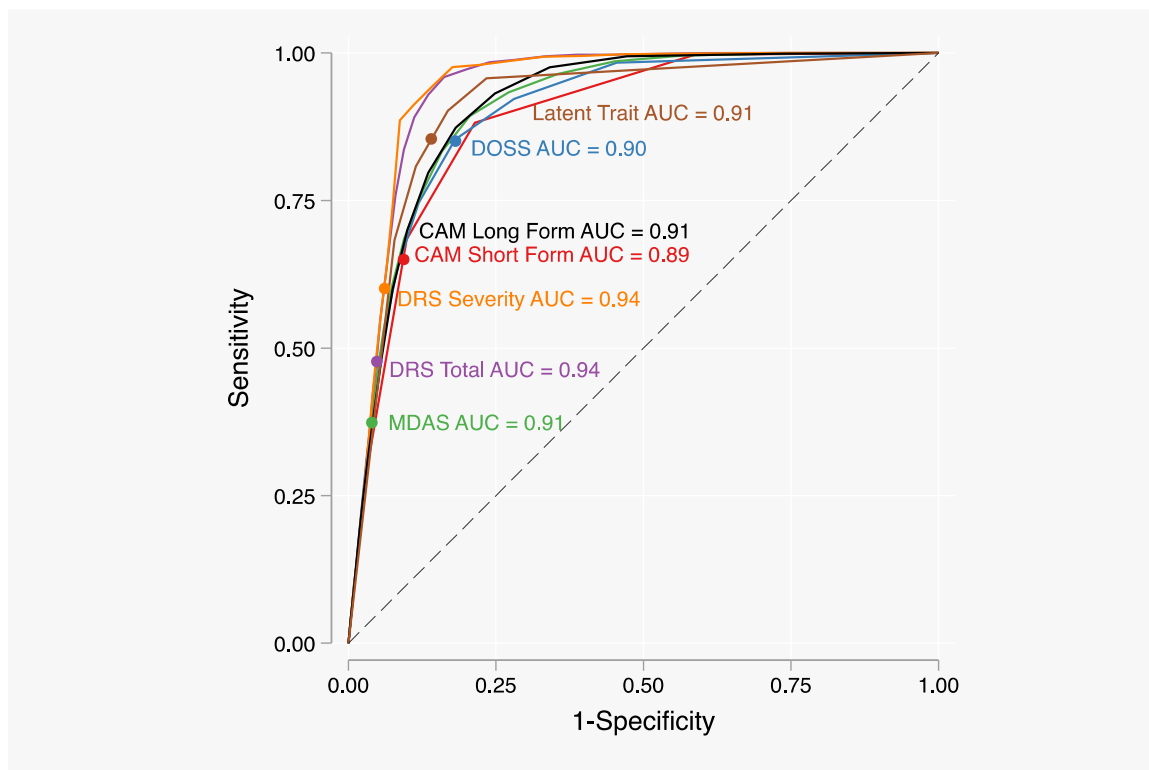
	Adamis et al. (N=200)	BASIL (N=352)	Detroyer et al. (N=48)
Age, years, mean (SD) or median (IQR)	81.1 (6.5)	80.3 (6.8)	72 (67.25; 78)
Sex:			
Female sex, n (%)	100 (50)	203 (58)	18 (38)
Male sex, n (%)	100 (50)	149 (42)	30 (62)
Race:			
White race, n (%)	NR	304 (86)	NR
Non-white race, n (%)	NR	48 (14)	NR
Years of education, mean (SD)	NR	14.5 (3.0)	NR
Married, n (%)	NR	139 (40)	26 (54)
Lives alone, n (%)	NR	135 (39)	7 (15)
Lives in nursing home, n (%)	NR	13 (3.7)	1 (2.1)
Dementia status:			
Dementia or history of cognitive impairment, n (%)	126 (63)	101 (29)	NR
No dementia or history of cognitive impairment, n (%)	74 (37)	251 (71)	NR
CAM delirium (ever), n (%)	34 (17)	88 (25)	11 (23)
DSM-5-defined delirium diagnosis, n (%)	26 (13)	NR	NR

NR = not reported; SD = standard deviation; IQR = interquartile range; BASIL = Better Assessment of Illness Study; CAM = confusion assessment method; DSM-5 = Diagnostic and Statistical Manual of Mental Disorders, 5th Edition

**Figure 4.1** shows ROC curves for each of the six delirium identification instruments and for the latent trait, propensity to delirium, using DSM-5 criteria for delirium as the reference standard. The area under each curve (AUC) ranged from 0.89-0.94. The dot on each curve represents the published author-described cut-point for that particular instrument. The DOSS is the only instrument where the author described cut-point occurs at Youden's J statistic, which is considered the optimal cut-point to simultaneously maximize sensitivity and specificity. All the other author described cut-points appeared to prioritize specificity over sensitivity. This is further demonstrated in **Table 4.2**, which shows for each instrument the cut-point nearest to 90% sensitivity, the cut-point nearest to 90% specificity, the cut-point that maximizes Youden's J statistic, and the author-described cut-point. The table presents the sensitivity and specificity at each cut-point. Additionally, the level on the latent trait, propensity to delirium, for each cut-point is shown. The latent trait is a continuous metric presumed to have a mean of 0 and standard deviation of 1. Importantly, we demonstrated that the latent trait level of 1 is the location on the metric that best describes case identification of delirium, since it yielded the cut-point that maximizes Youden's J statistic, with sensitivity of 91% and specificity 83%.



**Figure 4.1. ROC curves for each delirium identification instrument compared to DSM-5 criteria**



CAM=Confusion Assessment Method, DOSS=Delirium Observation Screening Scale, DRS-R-98=Delirium Rating Scale-Revised-98, MDAS=Memorial Delirium Assessment Scale; AUC=area under the curve

Figure legend: The receiver operating characteristic (ROC) curve for each of the instruments and their different scorings, plus the latent trait (propensity to delirium) is shown. Each curve displays the instrument AUC. The large dot on each curve is the author described cut-point on each instrument (except for the latent trait curve where the dot is Youden's J statistic). The CAM short form and long form each use the same diagnostic algorithm to identify delirium.

**Table 4.2. Instrument cut-points**

	CAM Short Form	CAM Long Form	DOSS	MDAS	DRS-R-98 Total	DRS-R-98 Severity	Latent Trait
<b>Cut-point nearest to 90% sensitivity</b>							
Cut-Point	2	5	2	6	12	10	1
Sensitivity	89%	87%	92%	89%	89%	90%	91%
Specificity	79%	82%	72%	79%	89%	89%	83%
Latent Trait	0.86	0.94	0.56	0.80	1.17	1.15	--
<b>Cut-point nearest to 90% specificity</b>							
Cut-Point	3	7	5	9	13	10	1.25
Sensitivity	69%	70%	63%	69%	84%	90%	81%
Specificity	90%	90%	92%	91%	91%	89%	89%
Latent Trait	1.55	1.36	1.52	1.38	1.31	1.15	--
<b>Cut-point nearest to Youden's J-Statistic</b>							
Cut-Point	2	5	3	6	11	8	1
Sensitivity	89%	87%	85%	89%	93%	97%	91%
Specificity	79%	82%	82%	79%	87%	82%	83%
Latent Trait	0.86	0.94	0.98	0.80	1.05	0.89	--
<b>Author described cut-point</b>							
Cut-Point	--	--	3	13	17.75	15.25	--
Sensitivity	48%	48%	85%	27%	32%	31%	--
Specificity	91%	91%	82%	96%	95%	95%	--
Latent Trait	--	--	0.98	1.99	1.82	1.85	--

CAM=Confusion Assessment Method, DOSS=Delirium Observation Screening Scale, DRS-R-98=Delirium Rating Scale-Revised-98, MDAS=Memorial Delirium Assessment Scale

In **Table 4.3**, each of the items across the four new example instruments is shown along with the source instrument of each item and the information of each item in descending order by information criteria. Example instrument 1 is a proposed short form that considered content validity, while example instrument 2 did not, having unbalanced domain content. Example instrument 3 is a proposed long form that considered content validity and included one item per delirium identification domain. Example instrument 4 contains 10 items selected only on the basis of information at a latent trait level of 1 and without regard to content balancing across domains. If clinical utility gives primary consideration to instrument length, example instruments 1 and 2 would be favored. Thus, example instruments 1 and 2 (short forms) would be more appropriate for use in a clinical setting where rapid assessment is needed, and intended to be followed by more in-depth diagnostic assessment for confirmation. In situations where reliability or accuracy were the primary consideration, example instruments 3 and 4 would be favored. Example instruments 3 and 4 could be best used for research purposes, a single stage diagnostic assessment, or more in-depth clinical interviews.

**Table 4.3. Four example instruments from the DEL-IB (Delirium Item Bank), each ordered by highest information**

Domain(s)/Item	Source Instrument	Information at latent trait level of 1
<b>EXAMPLE INSTRUMENT #1 – short form with content validity</b>		
<b>Disorganized Thinking:</b> rambling, irrelevant, or incoherent speech, or by tangential, circumstantial, or faulty reasoning	MDAS	18.7
<b>Inattention:</b> verbal perseverations, distractibility, and difficulty with set shifting	DRS-R-98	2.5
<b>Acute onset:</b> acute change in mental status from baseline	CAM	1.6
<b>Disorientation and Cognitive Impairment:</b> thinking he/she was somewhere other than the hospital, using the wrong bed, or misjudging the time of day	CAM	1.4
<b>Fluctuating Course:</b> symptoms come and go or increase and decrease in severity	CAM	0.5
<b>EXAMPLE INSTRUMENT #2 – short form with highest information</b>		
<b>Disorganized Thinking:</b> rambling, irrelevant, or incoherent speech, or by tangential, circumstantial, or faulty reasoning.	MDAS	18.7
<b>Disorganized Thinking:</b> disorganized or incoherent, such as rambling or irrelevant conversation, unclear or illogical flow of ideas, or unpredictable switching from subject to subject	CAM	5.6
<b>Disorganized Thinking:</b> abnormalities of thinking processes based on verbal or written output.	DRS-R-98	5.6
<b>Inattention:</b> verbal perseverations, distractibility, and difficulty with set shifting	DRS-R-98	2.5
<b>Inattention:</b> questions needing to be rephrased and/or repeated because patient's attention wanders, patient loses track, patient is distracted by outside stimuli or over-absorbed in a task.	MDAS	2.2

**EXAMPLE INSTRUMENT #3 – long form with content validity**

<b>Disorganized Thinking:</b> rambling, irrelevant, or incoherent speech, or by tangential, circumstantial, or faulty reasoning.	MDAS	18.7
<b>Inattention:</b> verbal perseverations, distractibility, and difficulty with set shifting difficulty	DRS-R-98	2.5
<b>Acute onset:</b> acute change in mental status from baseline	CAM	1.6
<b>Disorientation and Cognitive Impairment:</b> thinking he/she was somewhere other than the hospital, using the wrong bed, or misjudging the time of day	CAM	1.4
<b>Hallucinations, perceptual disorder, or distortion:</b> hallucinations, illusions, or misinterpretations	CAM	0.99
<b>Cognitive Impairment:</b> Short-term memory deficits	DRS-R-98	0.85
<b>Fluctuating course:</b> Fluctuation of symptom severity - waxing and waning of an individual or cluster of symptom(s)	DRS-R-98	0.81
<b>Level of consciousness and Inattention:</b> current awareness of and interaction with the environment	MDAS	0.75
<b>Psychomotor agitation:</b> picking, disorderly, restless	DOSS	0.44
<b>Psychomotor retardation:</b> reacts slowly to instructions	DOSS	0.22

**EXAMPLE INSTRUMENT #4 – long form with highest information**

<b>Disorganized Thinking:</b> rambling, irrelevant, or incoherent speech, or by tangential, circumstantial, or faulty reasoning.	MDAS	18.7
<b>Disorganized Thinking:</b> disorganized or incoherent, such as rambling or irrelevant conversation, unclear or illogical flow of ideas, or unpredictable switching from subject to subject	CAM	5.6
<b>Disorganized Thinking:</b> abnormalities of thinking processes based on verbal or written output	DRS-R-98	5.6
<b>Inattention:</b> verbal perseverations, distractibility, and difficulty with set shifting	DRS-R-98	2.5

<b>Inattention:</b> questions needing to be rephrased and/or repeated because patient's attention wanders, patient loses track, patient is distracted by outside stimuli or over-absorbed in a task	MDAS	2.2
<b>Inattention:</b> Maintains attention to conversation or action	DOSS	1.9
<b>Acute onset:</b> acute change in mental status from baseline	CAM	1.6
<b>Acute onset:</b> acuteness of onset of the initial symptoms of the disorder or episode being currently assessed	DRS-R-98	1.4
<b>Disorientation and Cognitive Impairment:</b> thinking he/she was somewhere other than the hospital, using the wrong bed, or misjudging the time of day	CAM	1.4
<b>Inattention and Disorganized thinking:</b> Does not finish question or answer	DOSS	1.1

CAM=Confusion Assessment Method, DOSS=Delirium Observation Screening Scale, DRS-R-98=Delirium Rating Scale-Revised-98, MDAS=Memorial Delirium Assessment Scale

**Table 4.4** shows sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) across each of the proposed example instruments. We show three cut-points for each instrument that could be used for different situations: screening, confirmation of diagnosis, and balanced high accuracy. The screening cut-point sought to maximize sensitivity (nearest to 90%), while the confirmation of diagnosis cut-point maximized specificity (nearest to 90%), and the balanced high accuracy cut-point was at the level of the latent trait that maximizes Youden's J statistic, each again with DSM-5 diagnostic criteria as the reference standard. Notably, the cut-point that optimized sensitivity and Youden's J statistic was the same across each of the example instruments. Also, of note, each cut-point across each instrument demonstrated a generally high NPV, while PPV was low. **Figure 4.2** shows ROC curves for each of the example instruments and for the latent trait, propensity to delirium, using DSM-5 criteria for delirium as the reference standard. The area under each curve (AUC) ranged from 0.91-0.92.

**Table 4.4. Psychometric properties of proposed new instruments**

New instruments derived from DEL-IB	No. of Items	Sensitivity	Specificity	PPV	NPV
<b><i>Example instrument #1 (short form with content validity – score range: 0-10, AUC=0.91)</i></b>					
Clinical Screening – optimize sensitivity (cut-point 3)	5	90%	81%	30%	99%
Clinical Confirmation of Diagnosis – optimize specificity (cut-point 5)	5	66%	91%	40%	97%
Clinical Balanced High Accuracy – Youden’s J statistic (cut-point 3)	5	90%	81%	30%	99%
<b><i>Example instrument #2 (short form with highest information – score range: 0-13, AUC=0.92)</i></b>					
Clinical Screening – optimize sensitivity (cut-point 3)	5	92%	79%	29%	99%
Clinical Confirmation of Diagnosis – optimize specificity (cut-point 6)	5	71%	91%	41%	98%
Clinical Balanced High Accuracy – Youden’s J statistic (cut-point 3)	5	92%	79%	29%	99%
<b><i>Example instrument #3 (long form with content validity – score range: 0-21, AUC=0.92)</i></b>					
Research Screening – optimize sensitivity (cut-point 6)	10	89%	83%	31%	99%
Research Confirmation of Diagnosis – optimize specificity (cut-point 8)	10	76%	89%	39%	98%
Research Balanced High Accuracy – Youden’s J statistic (cut-point 6)	10	89%	83%	31%	99%
<b><i>Example instrument #4 (long form with highest information – score range: 0-21, AUC=0.92)</i></b>					
Research Screening – optimize sensitivity (cut-point 6)	10	89%	82%	32%	99%
Research Confirmation of Diagnosis – optimize specificity (cut-point 9)	10	75%	89%	39%	98%



Research Balanced High Accuracy – Youden's J statistic (cut-point 6)	10	89%	82%	32%	99%
--	----	-----	-----	-----	-----

DEL-IB=Delirium Item Bank, PPV=positive predictive value; NPV=negative predictive value;  
AUC=area under the curve

**Figure 4.2. ROC curves for each example instrument compared to DSM-5 criteria**

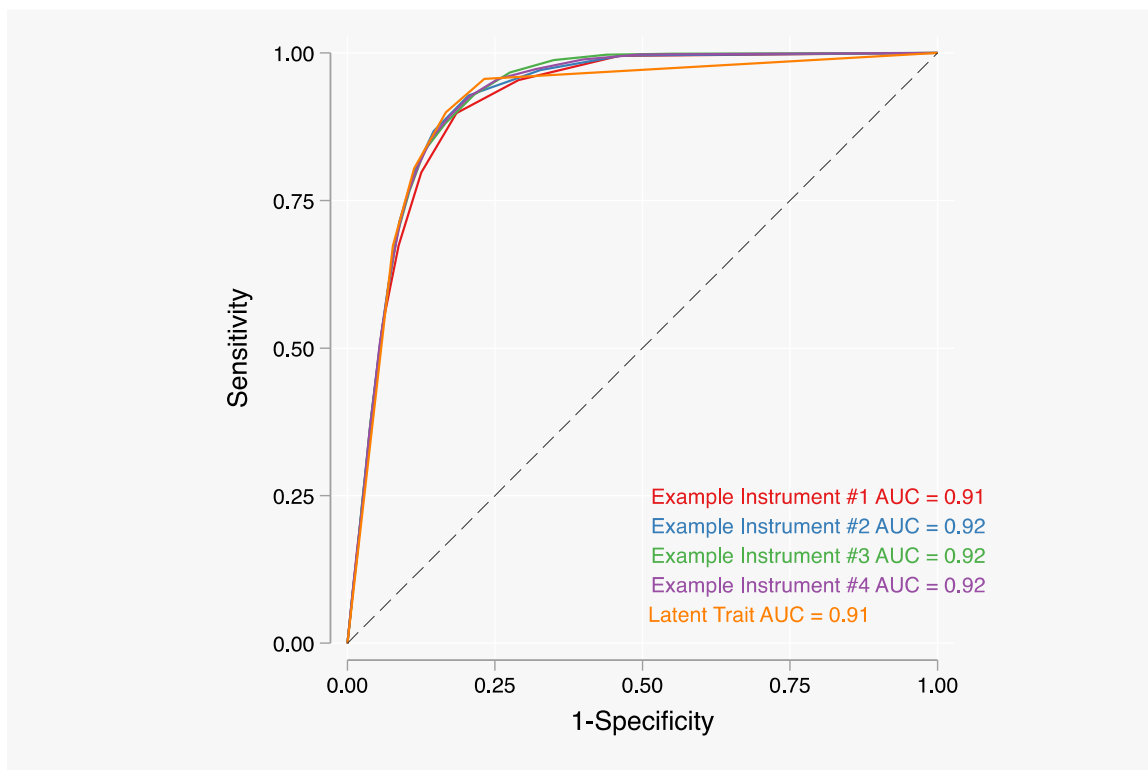


Figure legend: The receiver operating characteristic (ROC) curve for each of the example instruments, plus the latent trait (propensity to delirium) is shown. Each curve displays the instrument Area Under the Curve (AUC).

## ***Discussion***

This study provides a demonstration of the applications of an item bank, the DEL-IB, developed using advanced psychometric methods. First, we used the DEL-IB, that included items from the CAM, DOSS, DRS-R-98, and MDAS, to compare performance characteristics of these instruments. Next, we used the DEL-IB to create four new example instruments. The development and use of an item bank is highly novel in the field of delirium measurement research. Item banks have been used in educational testing for decades, but only recently have been applied in the field of measurement in healthcare. One recent advance in healthcare has been the use of modern methods in psychometrics to produce an item bank to fuel better measurement for patient-reported outcomes through the PROMIS (Patient-Reported Outcomes Measurement Information System) initiative (48).

In this study, we identified potential cut-points that optimized either sensitivity or specificity, or for balancing both sensitivity and specificity simultaneously. Interestingly, in general the cut-point that each author had originally chosen for their instrument tended to fall far from the balanced cut-point chosen on the basis of Youden's J statistic. The author-defined cut-points tended to have high specificity at the expense of sensitivity in our simulation.

Next, we used the DEL-IB to create new instruments and to evaluate their psychometric properties. While we had previously used the DEL-IB to create crosswalks between instruments [See Chapter III], the current work is another important demonstration of the usefulness of the DEL-IB. We displayed four example instruments, two short forms and two long forms, with one chosen to maximize content validity and the other to maximize psychometric information, respectively. We also recommended three separate cut-points that could be used on each instrument for different clinical or research purposes. Different customized instruments can be created to optimize clinical use across specialized settings and needs. For instance, screening tests would ideally be short-forms with high sensitivity; while diagnostic tests may be longer forms with high specificity. In settings with high prevalence of delirium, such as the intensive care unit, instruments with balanced accuracy may be preferred to minimize both false positives and false negatives. In the current study, it is key to note that while each cut-point had a high negative predictive value, they all had quite low positive predictive values. This means that if a participant were to test negative on any of the example instruments at any cut-point used, one could feel assured that they did not have delirium. However, if a participant were to test positive on any of the example instruments, further clinical evaluation would be required at any cut-point to confirm the diagnosis. Thus, these examples help to demonstrate how new psychometrically-based instruments can be developed using the DEL-IB.

The major strength of this study is that this is the only existing item bank for delirium identification instruments. The DEL-IB includes four different instruments, with a total of six different scoring methods, and 50 delirium assessment items. The DEL-IB was built from three international databases, which enhances generalizability. Another strength was the use of DSM-5-defined delirium diagnostic criteria as our reference standard, which is widely accepted as the current reference standard to evaluate the performance characteristics of each instrument. A further strength of this study stems from our previous expert panel work, which assessed each item and domain in DEL-IB for their content validity in delirium identification (8). This allowed us to easily select items across each domain vital to identifying delirium and the creation of example instruments that would uphold content validity.

There are several limitations that deserve comment. First, the Adamis et al. dataset had a delirium prevalence that was lower than the other two studies. This is important to note since our simulations were based on extrapolating results of the Adamis et al. study to the dataset. However, we performed direct adjustment for this prevalence difference in our models, so this effect was minimized. While we are not aware of any problems, we must acknowledge that any potential errors or idiosyncratic features of the diagnostic procedure from Adamis et al. will be propagated into our simulation results. Second, we used DSM-5-defined

delirium diagnostic criteria as our reference standard for the current study. However, it is essential to understand that the reference standard definition of delirium has evolved over time and will continue to do so, such that the instruments may perform differently based on which DSM reference standard was (or will be) used. Third, it should be noted that across the two example instruments that were created with only highest information considered, those examples are hindered by ‘bloated’ specific measurement. Bloated specific measurement refers to having instruments with too many items on a single domain of a construct, resulting in problems with content validity (27, 92). Fourth, the proposed instruments were developed as examples for using DEL-IB and are not intended for immediate clinical application. We did not consider the logistics of how to administer or order the items across the different example instruments. Refining and testing these instruments will be essential future work before these instruments—or any that are developed from the DEL-IB—are used in clinical practice. Fifth, it is a known problem in IRT that discrimination parameters can be biased upwards when maximum likelihood estimation procedures are used (93), as were used in this study. This can happen when trying to apply an IRT model to a set of items that are logically dependent upon one another, known as Guttman scales, e.g., difficulty carrying a 10-pound bag of groceries is logically dependent on difficulty carrying a 5-pound bag of groceries. This can also happen when items are *de facto* dependent upon one another because there is a relatively small sample size. The way to address this is to collect more data or

use a parameter estimation technique that places constraints on the allowable range of parameter estimates, such as Bayesian parameter estimation.

The creation of the DEL-IB is novel within the field of delirium research and has the potential to fundamentally advance the field. Based on our psychometric work, there is a potential case to be made that new cut-points may be appropriate on currently existing delirium identification instruments to aid in screening or diagnosing. Further investigation would be necessary to field test the proposed cut-points and new instruments in actual patient samples instead of simulated data. Field testing could also include examination of concurrent validity against DSM-5-defined delirium diagnosis as the reference standard and predictive validity against clinical outcomes. Additional next steps would include expanding on the DEL-IB by adding additional instruments from existing data sources with overlapping instruments. Ultimately, the goal is to find a single unified approach to identify delirium for the field and this work provides a fundamental step in that direction.

## ***Acknowledgment***

### Published Manuscript Co-Author Contributions

Mr. Helfand conceived of the project, collected the search, organized and convened the expert panel, synthesized expert panel feedback, did all analyses, wrote the manuscript, created all tables and figures. Mr. Helfand and Dr. Jones had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Other contributions include:

Concept and design: Helfand, Jones, Inouye, Boudreaux

Acquisition, analysis, or interpretation of data: Helfand, Jones, Boudreaux, Inouye, Detroyer, Milisen, Adamis.

Drafting of the manuscript: Helfand, Jones.

Critical revision of the manuscript for important intellectual content: Helfand, Metzger, Detroyer, Milisen, Adamis, Boudreaux, Inouye.

Statistical analysis: Helfand, Tommet, Jones.

Obtained funding: Helfand, Jones, Inouye.

Administrative, technical, or material support: Helfand, Jones, Inouye, Boudreaux.

Supervision: Helfand, Jones, Inouye, Boudreaux.



## CHAPTER V – Discussion and Future Directions

### ***Restatement of Specific Aims***

The overarching goal of this dissertation project was to apply advanced psychometric methods to improve the identification of delirium. This project proceeded with the following specific aims.

**Specific Aim 1.** Determine the 4 most commonly used and well-validated instruments for delirium identification through a systematic review of the medical literature, applying standardized methodologic quality ratings (Chapter II).

**Specific Aim 2.** Harmonize the 4 most commonly used and well-validated delirium assessment instruments to generate an item bank, which is a collection of the individual instrument questions or ratings along with their parameter estimates derived from item response theory (IRT) analyses (Chapter III).

**Specific Aim 3.** Explore applications of the harmonized item bank through several approaches. First, identifying different cut-points that will optimize: (a) balanced high accuracy (Youden's J-Statistic), (b) screening (sensitivity), and (c) confirmation of diagnosis (specificity) in identification of delirium. Second, comparing performance characteristics of short forms (versus long forms) developed from the item bank (Chapter IV).

### ***Summary of the Major Results***

In Chapter II, I reported on a systematic review and selection of high-quality delirium identification instruments. I conducted the systematic review by searching Cumulative Index to Nursing and Allied Health Literature (CINAHL), Cochrane Library, Excerpta Medica Database (Embase), PsycINFO, PubMed, and Web of Science from January 1, 1974, to January 31, 2020, with the keywords “delirium” and “instruments,” along with their known synonyms. I identified 2,542 articles potentially pertaining to delirium measurement, and of these 75 met eligibility criteria for detailed review. The eligibility criteria included English-language articles only and requiring the article to be a systematic review, meta-analysis, or narrative review that evaluated at least two different delirium identification instruments. I excluded studies restricted to alcohol-related delirium (delirium tremens) or pediatric populations, studies using only animal models, studies in which delirium was not the outcome, or not a review article. These articles referenced 30 different delirium identification instruments. Two reviewers assessed the eligibility of articles and extracted data on all potential delirium identification instruments. The original publication of each instrument underwent methodologic quality review of psychometric properties using Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) definitions. I convened a clinical expert panel that classified domains for delirium identification based on criteria from the Diagnostic and Statistical Manual of Mental Disorders (DSM)-III through DSM-5. I determined citation count through

Scopus for the original publication of each instrument. Then, I undertook a methodological quality review of psychometric properties for each instrument using COSMIN definitions. Four instruments were noteworthy for having at least two of three of the following: citation count of 200 or more, strong validation methodology in their original publication, and fulfillment of DSM-5 criteria. These were, alphabetically, the Confusion Assessment Method (CAM), Delirium Observation Screening Scale (DOSS), Delirium Rating Scale-Revised-98 (DRS-R-98), and Memorial Delirium Assessment Scale (MDAS).

In Chapter III, I reported on the statistical harmonization of the four selected instruments identified in my systematic review (above). This chapter involved a secondary data analysis from three studies, and a simulation study based on the observed data. I obtained data from three previous studies of hospitalized older adults (age 65+ years) in the United States, Ireland, and Belgium. One of these studies (Ireland) (72), included reference standard diagnoses according to DSM-5 criteria. The combined dataset included 600 participants, contributing 1,623 delirium assessments. Each of the studies that generated the data assessed participants with multiple delirium identification instruments. Using item response theory (IRT), I linked scores across instruments, placing all four instruments and their separate scorings on the same metric (the propensity to delirium). Kappa statistics comparing agreement in delirium identification among the instruments ranged from 0.37-0.75, with the highest between the DRS-R-98 total score and

MDAS. After linking scores, I created a harmonized item bank, called the Delirium Item Bank (DEL-IB), consisting of 50 items. The DEL-IB permitted me to create six crosswalks, which allow straightforward calculation of equivalent scores across instruments.

In Chapter IV, I reported on applications of the DEL-IB to evaluate and create instruments, specifically to find optimal cut-points on the four instruments from Chapter III and to demonstrate use of the DEL-IB to create new instruments. I again utilized the combined international dataset of hospitalized older adults of 600 participants introduced in Chapter III. I began by evaluating published cut-points and establishing new cut-points (optimizing sensitivity or specificity, and Youden's J statistic) on the latent trait, propensity to delirium, based on DSM-5-defined delirium diagnosis from a reference standard collected in the Adamis et. al. dataset. For example, the cut-point on the MDAS at Youden's J statistic was at a sum score of 6 with 89% sensitivity and 79% specificity. Then, I further explored the DEL-IB to create four example instruments (two short forms and two long forms) and evaluated their performance characteristics. The four example instruments illustrate differences when priority is given to brevity versus fidelity (short versus long) and when priority is given to sensitivity versus specificity (correctly identifying disease among those who truly have disease versus correctly ruling out disease among those who truly are disease free). These different prototypical instruments reflect choices made for different applications.

For example, clinicians may look to optimize sensitivity for the sake of screening, optimize specificity for diagnosing (especially when the next therapeutic step is very invasive like a brain biopsy), or apply findings to high-risk settings (i.e., intensive care units) with balanced accuracy. A short form optimized for sensitivity would be useful for quick screening in clinical settings, while a long form (with superior performance characteristics) could be used for more in-depth clinical interviews or research purposes.

### ***Products of this Work***

There are several major products of my work. From Chapter II, I have helped update the field by describing and characterizing the different **delirium identification instruments** in active use. I undertook a rigorous approach to comparing each of the instruments, resulting in our recommendation of the CAM, DOSS, DRS-R-98, and MDAS. Not only are these instruments widely used and demonstrate strong psychometric properties, but they are also fairly distinct in their target users and settings. For instance, the CAM was designed for use by non-psychiatrist clinicians and trained-lay raters. The DOSS was created for use by floor nurses. The DRS-R-98 and MDAS are typically used by trained psychiatrists. Thus, this study allowed us to identify instruments that would serve a broad swath of diverse users and patients, across multiple settings. This diversity of the instrument users and settings provided important context for Chapter III, since a major goal was to harmonize instruments that would be

useful across the field. Chapter III produced the **Delirium Item Bank (DEL-IB)**, which yielded multiple applications and products. I created **crosswalks** that allowed for the direct comparison of scores across 6 different delirium identification instruments and scorings in real time. Currently, our team is developing a **Harmonization Shiny App** that will be accessed openly (without charge) through the NIH-funded Network for Investigation of Delirium: Unifying Scientists (NIDUS) website (94). The hope is that clinicians will use the app to compare scores at the time of care, and that researchers can use the app to compare and combine scores across patients and across studies in making group inferences. In Chapter IV, the DEL-IB was used to further understand the psychometric properties of selected delirium identification instruments from Chapter II. I offered different potential cut-points that optimized sensitivity, specificity or Youden's J Statistic. I also compared the results to the author described cut-points. Additionally, I utilized the DEL-IB to create example **short and long form instruments** that one could use to accurately and rapidly identify delirium. I again suggested different potential cut-points that optimized sensitivity, specificity or Youden's J Statistic, which one could need for different clinical or research circumstances.

### ***Major Conclusions of the Work***

There are several major conclusions and implications of this body of work. The overall goal of this dissertation was to apply advanced psychometric methods to

improve measurement in the field of delirium. To achieve this goal, I applied state-of-the-art approaches of IRT, utilized in other fields of measurement, such as educational and psychological testing. I started with a systematic review of the existing medical literature to identify all of the current instruments in active use for identification of delirium, and selected key measures (based on a priori criteria) for my measurement work. Subsequently, I applied IRT approaches to harmonize the key measures I identified, which allowed me to statistically place them on the same metric, a latent trait called a propensity to delirium. Next, the IRT approaches also allowed me to create a delirium item bank (DEL-IB), which is a set of items (features of delirium) along with their parameter estimates, as a resource for the field. I demonstrated how to use the DEL-IB to create new measures to achieve different clinical goals (i.e., screening with maximal sensitivity, diagnosis with maximal specificity, or application to high-risk settings with balanced accuracy). Thus, this body of work provides the tools and methods to advance measurement in the field of delirium. In addition to the ability to create optimized measures for different settings and for different uses, these advances will allow for combination of data bases with harmonized outcomes; meta-analytic studies; or generation of large data bases (such as for omics or machine learning studies). These new applications of my work hold substantial promise for the future of the field.

### ***Strengths of the Work***

Several strengths of this work deserve comment. I have updated the field of delirium by describing and characterizing the currently used delirium identification instruments and thoroughly investigating the psychometric properties of each instrument in its original publication. In Chapters II and IV, I convened an expert panel of interdisciplinary delirium clinical and research experts. This holds great value in assuring that the results of the work align with diagnostic criteria, and hold content and face validity. The development of the DEL-IB utilized a novel approach to the field of delirium by applying advanced psychometric methods to three independent datasets that each studied multiple and overlapping delirium identification instruments. A further strength is the fact that these datasets were multinational including older hospitalized patients from the United States, Ireland, and Belgium, which heightens the generalizability of the results. Importantly, the inclusion of DSM-5 reference standard ratings in the Adamis et al. dataset, which is considered the current reference standard for delirium identification, assisted in evaluating the instruments' performance characteristics. Moreover, this is the first study in the field that I am aware of to use modern methods in psychometrics to harmonize 6 separate delirium identification instruments and their different scorings. The result of this work was the DEL-IB that is the only item bank in existence to date for delirium identification instruments.



### ***Caveats of the Conclusions***

There are caveats of the conclusions that require further discussion. In Chapter II, I utilized a systematic review of systematic reviews, which is an accepted approach. However, it is always possible that I may have missed some eligible studies and instruments. This is unlikely given the approach of reviewing the citations of the included articles and garnering input from experts in the field of delirium research. Restricting the psychometric review to the original publication of each instrument is another limitation, since it is possible that further investigation of the literature for validation studies for each instrument may have resulted in stronger psychometric evidence for each instrument. However, this would bias in favor of older instruments that have been in existence for a longer time, and therefore, may have had more validation studies published over time. The use of citation count could also bias towards older instruments, but this was only one of the selection criteria used by the expert panel. For the COSMIN rating, I only assessed the presence or absence of each of the validity or reliability criteria. The rankings of the instruments may have been different had I incorporated the actual performance statistics from the original publications. However, I decided not to use this approach since the studies used different reference standards, reported varying performance statistics, and examined disparate study samples of patients across diverse clinical settings. Additionally, since many patients in the intensive care unit (ICU) are non-verbal and the questions asked on those assessments are quite different, I decided to exclude

delirium identification instruments specific to the ICU setting. Unfortunately, this limits our ability to generalize my findings to the ICU setting. Further, I gave little consideration to distinguishing delirium in persons with underlying dementia. This is an aspect of the field of extreme importance that requires added future investigation. In future work, there will be a need to rank and rate delirium instruments for their ability to identify delirium superimposed on dementia or differentiate delirium and dementia.

Another caveat is the lack of any primary data collection for the work in Chapters III and IV. Moreover, the existing secondary data sources may not have applied the instruments consistently or coded them the same way across all sites. Each of the samples came from hospitalized older patients, thus, the results may not be generalizable to non-hospital settings for delirium identification. The Better Assessment of Illness (BASIL) study is the largest known study of multiple delirium instruments applied to each participant. However, even after combining each of the three datasets together for a total of 600 participants, the overall sample size was limited for this type of work, and not every response category was seen across each instrument within participants. Thus, I utilized simulation methods to help draw our conclusions. I based the simulations on extrapolating results of the Adamis et al. dataset, which had a lower delirium prevalence than the other two studies. I accounted for this by performing direct adjustment for the prevalence difference in the models to minimize these effects. Thus, another

potential limitation is that any inaccuracies in the assessment of DSM-5-defined delirium diagnosis procedures from Adamis et al. will be propagated into our simulation results. While I am not aware of any problems, I was not involved to assure high quality in the collection of the reference standard rating. A final limitation to mention is that the proposed example instruments developed from the DEL-IB are not ready for immediate clinical use at this time. There was little consideration given to the feasibility or logistics of administration of the instruments, or ordering of the items across the example instruments. The next steps would include refining and testing these instruments (or any other potential instruments developed from the DEL-IB) in a field study of patients to assure their feasibility and validity before application in clinical practice.

### ***Implications of the Work and Future Directions***

There are many implications of this work and future directions that research in the field of delirium should take. The major implication of this work leads directly to help interpret and combine current delirium studies, and help with developing new measures using the DEL-IB. In a future step, I could continue to expand the DEL-IB with additional existing studies that have another delirium instrument and one of CAM, DOSS, MDAS, DRS-R-98 or studies that have one of these instruments to help with the calibration by adding a greater sample size (i.e., a study with 1000 participants all only assessed by the DOSS). Currently, the NIDUS Research Hub lists over 600 different studies of delirium that could be

utilized to help expand the DEL-IB, and ultimately combined for meta-analyses or integrative analyses of pooled data. Another future direction to undertake would include validation of the derived example short and long forms from Chapter IV in a prospective multicenter cohort study. I could also mount integrative data analysis using one of four instruments to have directly comparable outcome measures across studies. Another major step includes the creation of big datasets. These could be synthesized via meta-analyses or integrative analysis for many potential uses. For example, the synthesis of results from multiple clinical trials could directly inform treatment recommendations, and assist with development of clinical guidelines and clinical practice standards. This kind of work has already been seen in the field of delirium with a recent systematic review on the risks of antipsychotic use for delirium, where outcomes were utilized across studies to draw conclusions and make recommendations (95). “Big data” is also needed for omics studies (e.g., genomics, proteomics, metabolomics, etc.) that advance the understanding of pathophysiology, which currently is poorly understood in delirium. Big data can further be applied to population-based prediction models that require large datasets, such as, machine learning or advanced prediction approaches. Item banks have been used for computerized adaptive testing (CAT), which enables the development of streamlined approaches for diagnosis (96). With a large enough item bank, and with enough participants, someday it may be possible that the DEL-IB could be used to create a CAT that could rapidly and accurately identify delirium across all

settings. Ultimately, this dissertation lays the groundwork for many future directions in the field of delirium research and clinical care.

### ***Unexpected Results and Personal Reflection***

There were a number of unexpected results and important lessons that I learned along this journey. I learned how to perform and finish a systematic review from start to finish, which turned out to be much more time-consuming than I envisioned in the beginning of my thesis work. I learned the value of collaborating with other groups with similar interests, who were able and willing to share data. However, the task of understanding the different data sets and sorting them was arduous and intensive to find the proper data for the needs of this dissertation work. I gained expertise in measurement, psychometrics, and test development, which on top of my pre-existing knowledge of epidemiology are skills that I now realize will ultimately translate to any field I decide to pursue. I did not fully grasp all of the potential uses for an item bank and the large number of applications that have real-world practical value, i.e., harmonization, crosswalks, and creation of new instruments. The DEL-IB could be enriched further with a large number of additional items across all delirium identification instruments. This was the first study I undertook using simulation methods. I discovered how I can use simulation to help solve many complex problems and can extrapolate to many other situations. I learned to be flexible in my scientific approach and managing expectations, such as adjusting to new results or findings that are uncovered

unexpectedly. These unexpected discoveries often lead to a need to adapt and adjust one's thoughts and timeline. Additionally, the COVID-19 pandemic changed the entire global landscape and my day-to-day work on my dissertation, and I had to adapt and persevere. I have also learned about grant writing and the entire National Institutes of Health (NIH) application process by writing an R36 grant to support my dissertation work, which was successfully funded.

### ***Influence on Future Career Directions***

I believe the work on this dissertation will influence and impact my future career in many tangible and intangible ways. I have truly gained an appreciation for rigorous scientific discovery, and I see myself as an academic physician scientist moving forward. I have learned how to think critically about evidence and how to problem solve skillfully. The field of measurement, epidemiology, and public health will pertain to any biomedical field and I see myself continuing in related research, applying measurement techniques I have learned to other topics in the field of medicine I choose for my career. As for the next steps in my career, I am planning to pursue a residency in internal medicine. While I do not yet know what specific fellowship I would wish to pursue, I do know I will continue to hold a strong interest in aging, epidemiologic, and measurement research, and I hope to build on these areas and strengths moving forward. I am confident that this dissertation work has laid a solid foundation for any future direction that I pursue.

## BIBLIOGRAPHY

1. Inouye SK, van Dyck CH, Alessi CA, Balkin S, Siegel AP, Horwitz RI. Clarifying confusion: the confusion assessment method: a new method for detection of delirium. *Annals of internal medicine*. 1990;113(12):941-8.
2. Ely EW, Inouye SK, Bernard GR, Gordon S, Francis J, May L, Truman B, Speroff T, Gautam S, Margolin R. Delirium in mechanically ventilated patients: validity and reliability of the confusion assessment method for the intensive care unit (CAM-ICU). *JAMA*. 2001;286(21):2703-10.
3. Ely EW, Margolin R, Francis J, May L, Truman B, Dittus R, Speroff T, Gautam S, Bernard GR, Inouye SK. Evaluation of delirium in critically ill patients: validation of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU). *Critical Care Medicine*. 2001;29(7):1370-9.
4. Inouye SK, Kosar CM, Tommet D, Schmitt EM, Puelle MR, Saczynski JS, Marcantonio ER, Jones RN. The CAM-S: development and validation of a new scoring system for delirium severity in 2 cohorts. *Annals of internal medicine*. 2014;160(8):526-33. doi: 10.7326/M13-1927. PubMed PMID: 24733193.
5. Inouye SK, Westendorp RGJ, Saczynski JS. Delirium in elderly people. *The Lancet*. 2014;383(9920):911-22. doi: 10.1016/S0140-6736(13)60688-1.
6. Fong TG, Tulebaev SR, Inouye SK. Delirium in elderly adults: diagnosis, prevention and treatment. *Nature Reviews Neurology*. 2009;5(4):210-20. doi: 10.1038/nrneurol.2009.24.
7. Chadwick J, Mann WN. *The medical works of Hippocrates*: Oxford Blackwell Scientific Publications; 1950.
8. Helfand BKI, D'Aquila ML, Tabloski P, Erickson K, Yue J, Fong TG, Hshieh TT, Metzger ED, Schmitt EM, Boudreaux ED, Inouye SK, Jones RN. Detecting

Delirium: A Systematic Review of Identification Instruments for Non-ICU Settings. *J Am Geriatr Soc.* 2020. Epub 2020/11/03. doi: 10.1111/jgs.16879. PubMed PMID: 33135780.

9. Jones RN, Cizginer S, Pavlech L, Albuquerque A, Daiello LA, Dharmarajan K, Gleason LJ, Helfand B, Massimo L, Oh E. Assessment of instruments for measurement of delirium severity: a systematic review. *JAMA Internal Medicine.* 2019;179(2):231-9.

10. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders.* Spitzer R, editor. Washington, DC. 1980.

11. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders.* 5th ed. Washington, DC: American Psychiatric Society; 2013.

12. MacLulich AM, Hall RJ. *Who understands delirium? : Oxford University Press;* 2011. p. 412-4.

13. Breitbart W, Gibson C, Tremblay A. The delirium experience: delirium recall and delirium-related distress in hospitalized patients with cancer, their spouses/caregivers, and their nurses. *Psychosomatics.* 2002;43(3):183-94.

14. Cole MG. Delirium in elderly patients. *American Journal of Geriatric Psychiatry.* 2004;12(1):7-21. doi: 10.1176/appi.ajgp.12.1.7. PubMed PMID: WOS:000188001000002.

15. Witlox J, Eurelings LS, de Jonghe JF, Kalisvaart KJ, Eikelenboom P, van Gool WA. Delirium in elderly patients and the risk of postdischarge mortality, institutionalization, and dementia: a meta-analysis. *JAMA.* 2010;304(4):443-51. Epub 2010/07/29. doi: 10.1001/jama.2010.1013. PubMed PMID: 20664045.



16. Oh ES, Fong TG, Hshieh TT, Inouye SK. Delirium in older persons: Advances in diagnosis and treatment. *JAMA*. 2017;318(12):1161-74. doi: 10.1001/jama.2017.12067.
17. Leslie DL, Marcantonio ER, Zhang Y, Leo-Summers L, Inouye SK. One-year health care costs associated with delirium in the elderly population. *Archives of internal medicine*. 2008;168(1):27-32.
18. Inouye SK, Bogardus Jr ST, Charpentier PA, Leo-Summers L, Acampora D, Holford TR, Cooney Jr LM. A multicomponent intervention to prevent delirium in hospitalized older patients. *New England journal of medicine*. 1999;340(9):669-76.
19. de la Cruz M, Fan J, Yennu S, Tanco K, Shin S, Wu J, Liu D, Bruera E. The frequency of missed delirium in patients referred to palliative care in a comprehensive cancer center. *Supportive Care in Cancer*. 2015;23(8):2427-33.
20. Kennedy M, Helfand BK, Gou RY, Gartaganis SL, Webb M, Moccia JM, Bruursema SN, Dokic B, McCulloch B, Ring H, Margolin JD, Zhang E, Anderson R, Babine RL, Hshieh T, Wong AH, Taylor RA, Davenport K, Teresi B, Fong TG, Inouye SK. Delirium in Older Patients With COVID-19 Presenting to the Emergency Department. *JAMA Network Open*. 2020;3(11):1-12.
21. Garcez FB, Aliberti MJ, Poco PC, Hiratsuka M, Takahashi SdF, Coelho VA, Salotto DB, Moreira ML, Jacob - Filho W, Avelino - Silva TJ. Delirium and adverse outcomes in hospitalized patients with COVID - 19. *Journal of the American Geriatrics Society*. 2020;68(11):2440-6.
22. Hand DJ. *Measurement: A Very Short Introduction*. Oxford, UK: Oxford University Press; 2016. 127 p.

23. Stevens SS. On the theory of scales of measurement. *Science*. 1946;103(2684):677-80.
24. Breitbart W, Rosenfeld B, Roth A, Smith MJ, Cohen K, Passik S. The Memorial Delirium Assessment Scale. *Journal of pain and symptom management*. 1997;13(3):128-37.
25. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. COSMIN checklist manual. Amsterdam: University Medical Center. 2012.
26. Terwee CB, B ML, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research*. 2012;4(1573-2649 (Electronic)):651-7.
27. De Vet HC, Terwee CB, Mokkink LB, Knol DL. *Measurement in medicine: a practical guide*: Cambridge University Press; 2011.
28. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychological bulletin*. 1955;52(4):281-302.
29. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*. 1995;50(9):741.
30. Schotte C, Maes M, Cluydts R, De Doncker D, Cosyns P. Construct validity of the Beck Depression Inventory in a depressive population. *Journal of Affective Disorders*. 1997;46(2):115-25.
31. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis*. 1985;38(1):27-36. Epub 1985/01/01. doi: 10.1016/0021-9681(85)90005-0. PubMed PMID: 3972947.

32. Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. *Journal of clinical epidemiology*. 1992;45(11):1201-18.
33. Streiner DL. Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of personality assessment*. 2003;81(3):209-19.
34. Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, Moons KG, Reitsma JB. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS Med*. 2013;10(10):e1001531. Epub 2013/10/22. doi: 10.1371/journal.pmed.1001531. PubMed PMID: 24143138; PMCID: PMC3797139.
35. Kimchi EY, Hshieh TT, Guo R, Wong B, O'Connor M, Marcantonio ER, Metzger ED, Strauss J, Arnold SE, Inouye SK. Consensus Approaches to Identify Incident Dementia in Cohort Studies: Systematic Review and Approach in the Successful Aging after Elective Surgery Study. *Journal of the American Medical Directors Association*. 2017;18(12):1010-8. e1.
36. Mariz J, Costa Castanho T, Teixeira J, Sousa N, Correia Santos N. Delirium Diagnostic and Screening Instruments in the Emergency Department: An Up-to-Date Systematic Review. *Geriatrics (Basel)*. 2016;1(3). Epub 2016/09/01. doi: 10.3390/geriatrics1030022. PubMed PMID: 31022815; PMCID: PMC6371145.
37. McCusker J, Cole MG, Voyer P, Ciampi A, Monette J, Champoux N, Vu M, Belzile E. Development of a delirium risk screening tool for long-term care facilities. *Int J Geriatr Psychiatry*. 2012;27(10):999-1007. Epub 2012/03/01. doi: 10.1002/gps.2812. PubMed PMID: 22367973.

38. Schuurmans MJ, Shortridge-Baggett LM, Duursma SA. The Delirium Observation Screening Scale: a screening instrument for delirium. *Research and theory for nursing practice*. 2003;17(1):31-50.
39. Lord FM. The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*. 1953;13(4):517-49.
40. Rasch G. *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. 1960.
41. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*. 1968.
42. Streiner DL, Norman GR, Cairney J. *Health measurement scales: a practical guide to their development and use*: Oxford University Press, USA; 2015.
43. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory*: Sage; 1991.
44. Jones RN. Differential item functioning and its relevance to epidemiology. *Curr Epidemiol Rep*. 2019;6:174-83. Epub 2019/12/17. doi: 10.1007/s40471-019-00194-5. PubMed PMID: 31840016; PMCID: PMC6910650.
45. Helfand M, Berg A, Flum D, Gabriel S, Normand S. Draft Methodology Report: "Our Questions, Our Decisions: Standards for Patient-centered Outcomes Research". Patient-Centered Outcomes Research Institute, 2012.
46. Dorans NJ. Linking scores from multiple health outcome instruments. *Quality of Life Research*. 2007;16(1):85-94.
47. Griffith L, van den Heuvel E, Fortier I, Hofer S, Raina P, Sohel N, Payette H, Wolfson C, Belleville S. *Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis*. Rockville, MD, USA: Agency

for Healthcare Research and Quality - AHRQ Methods for Effective Health Care; 2013. Available from: [www.effectivehealthcare.ahrq.gov/reports/final.cfm](http://www.effectivehealthcare.ahrq.gov/reports/final.cfm).

48. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B, Rose M. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Medical care*. 2007;45(5 Suppl 1):S3.
49. Adamis D, Sharma N, Whelan PJ, Macdonald AJ. Delirium scales: a review of current evidence. *Aging & mental health*. 2010;14(5):543-55.
50. Inouye SK, Foreman MD, Mion LC, Katz KH, Cooney Jr LM. Nurses' recognition of delirium and its symptoms: comparison of nurse and researcher ratings. *Archives of internal medicine*. 2001;161(20):2467-73.
51. Neufeld KJ, Nelliott A, Inouye SK, Ely EW, Bienvenu OJ, Lee HB, Needham DM. Delirium diagnosis methodology used in research: A survey-based study. *The American Journal of Geriatric Psychiatry*. 2014;22(12):1513-21. doi: 10.1016/j.jagp.2014.03.003. PubMed PMID: 2014-48897-020.
52. Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, Knol DL, Bouter LM, de Vet HC. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research*. 2009;18(3):313-33.
53. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*. 2009;6(7):e1000097.
54. Eden J, Levit L, Berg A, Morton S. Finding what works in health care: standards for systematic reviews: National Academies Press; 2011.

55. Peters JL, Sutton AJ, Jones DR, Rushton L, Abrams KR. A systematic review of systematic reviews and meta-analyses of animal experiments with guidelines for reporting. *Journal of Environmental Science and Health Part B: Pesticides, Food Contaminants, and Agricultural Wastes*. 2006;41(7):1245-58.
56. Smith V, Devane D, Begley CM, Clarke M. Methodology in conducting a systematic review of systematic reviews of healthcare interventions. *BMC medical research methodology*. 2011;11(1):15.
57. Morrison A, Polisena J, Husereau D, Moulton K, Clark M, Fiander M, Mierzwinski-Urban M, Clifford T, Hutton B, Rabb D. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. *International journal of technology assessment in health care*. 2012;28(2):138-44.
58. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*. 2016;5(1):210.
59. Gélinas C. Delirium Assessment Tools for Use in Critically Ill Adults: A Psychometric Analysis and Systematic Review. *Critical Care Nurse*. 2018;38(1):38-50. doi: 10.4037/ccn2018633. PubMed PMID: 127417951. Language: English. Entry Date: 20180127. Revision Date: 20180127. Publication Type: Article.
60. Dalkey N, Helmer O. An experimental application of the Delphi method to the use of experts. *Management science*. 1963;9(3):458-67.
61. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. Fourth Edition, Text Revision ed. Washington, DC: American Psychiatric Association; 2000.

62. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Third edition, Revised ed. Washington, DC: American Psychiatric Association; 1987.
63. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Fourth Edition ed. Washington, DC: American Psychiatric Association; 1994.
64. Schulman-Green D, Schmitt EM, Fong TG, Vasunilashorn SM, Gallagher J, Marcantonio ER, Brown CH, Clark D, Flaherty JH, Gleason A. Use of an expert panel to identify domains and indicators of delirium severity. *Quality of Life Research*. 2019;1-14.
65. Wei LA, Fearing MA, Sternberg EJ, Inouye SK. The Confusion Assessment Method: a systematic review of current usage. *J Am Geriatr Soc*. 2008;56(5):823-30. Epub 2008/04/04. doi: 10.1111/j.1532-5415.2008.01674.x. PubMed PMID: 18384586; PMCID: PMC2585541.
66. Shi Q, Warren L, Saposnik G, MacDermid JC. Confusion assessment method: A systematic review and meta-analysis of diagnostic accuracy. *Neuropsychiatric Disease and Treatment*. 2013;9. PubMed PMID: 2013-34061-001.
67. van Velthuisen EL, Zwakhalen SMG, Warnier RMJ, Mulder WJ, Verhey FRJ, Kempen G. Psychometric properties and feasibility of instruments for the detection of delirium in older hospitalized patients: a systematic review. *International Journal of Geriatric Psychiatry*. 2016;31(9):974-89. doi: 10.1002/gps.4441. PubMed PMID: WOS:000382959300002.
68. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM. QUADAS-2: a revised tool for the quality

assessment of diagnostic accuracy studies. *Annals of internal medicine*. 2011;155(8):529-36.

69. Marcantonio ER. Delirium in Hospitalized Older Adults. *N Engl J Med*. 2017;377(15):1456-66. Epub 2017/10/12. doi: 10.1056/NEJMcp1605501. PubMed PMID: 29020579; PMCID: PMC5706782.

70. Hshieh TT, Fong TG, Schmitt EM, Marcantonio ER, D'Aquila M, Gallagher J, Xu G, Gou YR, Abrantes TF, Bertrand SE, Jones RN, Inouye SK. The Better Assessment of Illness Study (BASIL) for Delirium Severity: Study design, procedures, and cohort description. *Gerontology*. 2018;65(1):20-9.

71. Detroyer E, Clement PM, Baeten N, Pennemans M, Decruyenaere M, Vandenberghe J, Menten J, Joosten E, Milisen K. Detection of delirium in palliative care unit patients: a prospective descriptive study of the Delirium Observation Screening Scale administered by bedside nurses. *Palliative Medicine*. 2014;28(1):79-86.

72. Adamis D, Rooney S, Meagher D, Mulligan O, McCarthy G. A comparison of delirium diagnosis in elderly medical inpatients using the CAM, DRS-R-98, DSM-IV and DSM-5 criteria. *International Psychogeriatrics*. 2015;27(6):883-9.

73. Inouye SK. *The Confusion Assessment Method (CAM): training manual and coding guide*. New Haven: Yale University School of Medicine. 2003;2(3):4.

74. Inouye SK. *The short confusion assessment method (short CAM): training manual and coding guide*. 2014.

75. Mutch WAC, El-Gabalawy R, Girling L, Kilborn K, Jacobsohn E. End-tidal hypocapnia under anesthesia predicts postoperative delirium. *Frontiers in neurology*. 2018;9:678.



76. Trzepacz PT, Mittal D, Torres R, Canary K, Norton J, Jimerson N. Validation of the Delirium Rating Scale-revised-98: comparison with the delirium rating scale and the cognitive test for delirium. *The Journal of neuropsychiatry and clinical neurosciences*. 2001;13(2):229-42.
77. Schuurmans MJ, Donders ART, Duursma SA, Shortridge-Baggett LM. Delirium case finding: pilot testing of a new screening scale for nurses. *Journal of the American Geriatric Society*. 2002;50(4):S3.
78. Schuurmans MJ, Deschamps PI, Markham SW, Shortridge-Baggett LM, Duursma SA. The measurement of delirium: review of scales. *Research and theory for nursing practice*. 2003;17(3):207-24. Epub 2003/12/06. PubMed PMID: 14655974.
79. Shih RA, Lee J, Das L. Harmonization of cross-national studies of aging to the health and retirement study cognition. RAND Corporation, 2012.
80. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969.
81. Choi SW. Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement*. 2009;33(8):644-5.
82. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *biometrics*. 1977:159-74.
83. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276-82. Epub 2012/10/25. PubMed PMID: 23092060; PMCID: PMC3900052.
84. Gross AL, Tommet D, D'Aquila M, Schmitt E, Marcantonio ER, Helfand B, Inouye SK, Jones RN, BASIL Study Group. Harmonization of delirium severity

instruments: a comparison of the DRS-R-98, MDAS, and CAM-S using item response theory. *BMC medical research methodology*. 2018;18(1):92.

85. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and optimal cut - point estimated from observations affected by a lower limit of detection. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*. 2008;50(3):419-30.

86. Naing NN. Easy way to learn standardization: direct and indirect methods. *The Malaysian journal of medical sciences: MJMS*. 2000;7(1):10.

87. Ly A, Marsman M, Verhagen J, Grasman RP, Wagenmakers E-J. A tutorial on Fisher information. *Journal of Mathematical Psychology*. 2017;80:40-55.

88. Baker FB. *The Basics of Item Response Theory*. Second ed. College Park, Maryland: ERIC Clearinghouse on Assessment and Evaluation; 2001.

89. Magis D. A note on the equivalence between observed and expected information functions with polytomous IRT models. *Journal of Educational and Behavioral Statistics*. 2015;40(1):96-105.

90. Baker FB, Kim S-H. *Item response theory: Parameter estimation techniques*: CRC Press; 2004.

91. Nunnally JC. *Psychometric Theory*. 3rd ed: Tata McGraw-Hill Education; 1994.

92. Cattell RB, Tsujioka B. The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*. 1964;24(1):3-30.

93. Agresti A. *An introduction to categorical data analysis*. Third ed. New York: John Wiley & Sons; 2019.

94. Fick DM, Auerbach AD, Avidan MS, Busby-Whitehead J, Ely EW, Jones RN, Marcantonio ER, Needham DM, Pandharipande P, Robinson TN, Schmitt EM, Trivison TG, Inouye SK. Network for Investigation of Delirium across the U.S.: Advancing the Field of Delirium with a New Interdisciplinary Research Network. *J Am Geriatr Soc.* 2017;65(10):2158-60. Epub 2017/06/21. doi: 10.1111/jgs.14942. PubMed PMID: 28631268; PMCID: PMC5641224.
95. Nikoosie R, Neufeld KJ, Oh ES, Wilson LM, Zhang A, Robinson KA, Needham DM. Antipsychotics for Treating Delirium in Hospitalized Adults: A Systematic Review. *Ann Intern Med.* 2019. Epub 2019/09/03. doi: 10.7326/m19-1860. PubMed PMID: 31476770.
96. Meijer RR, Nering ML. Computerized adaptive testing: Overview and introduction. *Applied psychological measurement.* 1999;23(3):187-94.