The measurement of collaboration

within healthcare settings

by

Stephen John Walters

A thesis submitted in partial fulfillment of the requirements for the degree

of

Master of Clinical Science

University of Adelaide

August, 2015

I certify that this work contains no material which has been accepted for the award of any other degree in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

_____

Stephen John Walters

ABSTRACT

The purpose of this study was to evaluate the measurement of collaboration within healthcare settings, with the aim of identifying validated instruments that measure collaboration in settings populated by a complex mix of participant types. To achieve this aim, a systematic review of measurement properties of instruments was performed following the Joanna Briggs Institute approach to systematic reviews and using the COSMIN checklist for methodological appraisal of validation studies.

A protocol for a systematic review was developed which established the criteria for inclusion of studies and defined the population to include more than two participant types. The focus of the review was the validation of instruments measuring collaboration, therefore validation studies were included. Clinical trials, observational studies and case studies were to be included where the study contributed to the interpretability of the instrument. Because the principal interest was healthcare, studies not about health or social care delivery were excluded. A search algorithm was developed and used search terms such as collaboration, interprofessional relations, psychometrics, measurement, reliability, instrument validation, factor analysis and instrument construction. Multiple databases were searched for published and unpublished studies.

As a result of the literature search and a refinement of the results, 21 studies of 12 unique instruments that met the inclusion criteria were included in methodological appraisal. Two appraisers reached consensus regarding the rating for methodological quality of the 21 studies and subsequently all were included in the review. The results were tabulated using a pre-established standard for this type of reporting. Tables for the characteristics of each

study accompany the results. A narrative synthesis was performed for the factor structures of the 12 instruments. This resulted in nine summary attributes that comprise collaboration; organizational settings, support structures, purpose and goals; communication; reflection on process; cooperation; coordination; role interdependence and partnership; relationships; newly created professional activities and professional flexibility.

From this process of rigorous analysis the author concluded that the measurement of social behavior like collaboration is problematic and traditional approaches to measurement using Classical Test Theory models may be limited. An approach to measurement of collaboration using Item Response Theory models should be considered. Furthermore, issues like measurement invariance and the limited use of triangulation methods in measurement and validation studies needs further research and development. An approach to measurement that incorporates an understanding of complexity and biopsychosocial principles presents a challenge for future research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

To all those who supported and encouraged me throughout the two years it took to complete this study, especially my supervisors, Dr. Suzanne Robertson-Malt and Dr. Cindy Stern who have provided expert guidance. Thank you.

I offer my appreciation to Janelle Jacobsen for co-appraising studies and to Dr. Catalin Tufanaru for advice on appraisal tools.

I dedicate this work to my late father and mother who always supported and encouraged me to become educated.

Stephen

GLOSSARY

Biopsychosocial: a term meaning to consider a persons' biological, psychological and social makeup as a way of viewing the human condition as a continuum of connected and nested hierarchies.[1]

Collaboration: occurs when a group of autonomous stakeholders of a problem domain engage in an interactive process, using shared rules, norms and structures, to act or decide on issues related to that domain.[2]

Complex Adaptive System: a collection of individual agents, who act freely in ways that are not always predictable and whose actions have an effect on other agents within the system.[3]

Complexity: incorporates a view of phenomena that considers the interconnectedness of elements and the importance of the environment in which the elements exist, known as a Complex Adaptive System.[4]

COSMIN: COnsensus on the **S**election of **M**easurement **IN**struments (COSMIN) is a methodological appraisal tool for assessing the measurement properties of instruments for the purpose of rating a measurement instrument's quality (validity, reliability and interpretability).

Evidence Based Healthcare: clinical decision-making that considers the best available evidence; the context in which the care is delivered; client preference; and the professional judgement of the health professional.[5]

Evidence synthesis: methodologies aimed at integrating multiple quantitative or qualitative data sets to determine the concordance and the magnitude of effect from multiple studies.[6]

Healthcare setting (HCS): **t**he HCS is any place where optimizing human health is the central activity of that setting and may include settings involved in the diagnosis and treatment of disease, the prevention of disease, the education of people to improve vitality and wellbeing, care of the elderly or disabled, palliation for people dying, and the rehabilitation of people with injury or post medical interventions.

Interpretability: the capacity of a metric produced by a measurement instrument to be translated to a qualitative meaning that is clinically or commonly understood.[7]

Reliability: a quantitative estimate of a measurement instrument's capacity to reproduce a metric within a specified tolerance for measurement error given similar or variable conditions for measurement or the degree to which the measurement is free from measurement error.[7]

Systematic review: a collation of all evidence that fits pre-specified eligibility criteria.[8]

Validation research/study: any scientific study reporting the results estimating the validity and reliability of a measurement instrument.

Validity: according to Messick,

> *"…an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment".*[9]

The aims of this thesis

This thesis aims to critically analyse the current state of measurement of collaboration within healthcare settings (HCS). The thesis presents the results of a systematic review, the purpose of which was to identify, appraise and rate measurement tools that quantify collaboration in HCS that have been validated with a sample that represents a complex mix of participant types.

The organisation of this thesis

The organisation of the thesis commences in Chapter 1 with a statement regarding the relevance of patient safety and presents an overview of the discourse within healthcare literature that positions collaboration as a key component of quality patient care. Definition of the HCS and the elements that comprise any HCS are presented including a description of various styles of team practice relevant to collaborative practice. Following is a definition of collaboration, the importance of measuring collaboration relative to teamwork and a brief overview of the central theories underpinning the measurement of collaboration within healthcare systems.

Chapter 2 addresses methodological principles upon which the systematic review process is based. This includes the science of evidence, evidence synthesis, systematic review, and measurement principles relevant to the concept of validity.

Chapter 3 details the method of systematic reviews of measurement properties of instruments and reproduces the systematic review protocol produced for the purpose of guiding the systematic review process. This protocol has been published in the Joanna Briggs Institute Library.

Chapter 4 presents the results of the systematic review and includes the search results, description of studies and the appraisal of methodological quality of each individual study. This is presented as a narrative synthesis.

Chapter 5 concludes the thesis with a discussion of the results of the systematic review. Issues relating to the measurement of complex, biopsychosocial phenomena are discussed and implications of the study results for clinical practice and further research are posited.

Chapter 1: Introduction

The primary focus of this chapter is to present the reader with the value of interdisciplinary collaboration in improving patient health and safety. A definition of collaboration is presented and a statement linking the value of measuring collaboration with teamwork outcomes is given. The chapter concludes with a summary of the theoretical basis of measuring collaboration.

1.1 Patient safety and the quality of healthcare systems

The Institute of Medicine in the United States published two reports in 2000 and 2001 addressing patient safety and quality management in healthcare.[10, 11] These reports heralded the need for change in the systems and process of care by highlighting that preventable adverse events and management errors in hospitals contributed to the death of people estimated to be upwards of 40,000 per year.[10] This made preventable medical error the cause of more deaths than each of motor vehicle accidents, breast cancer and AIDS. This report galvanised quality departments across HCS to critically review their systems and processes of care towards identifying areas for improvement. Healthcare has safety and quality problems because of outmoded systems of work.[11(p4)] A key recommendation from the report was the impact that collaborative teamwork can have on improving patient outcomes. The authors of one report mention interdisciplinary collaboration "...becom[ing] increasingly necessary for redesigning complex systems of care..." to address "...errors ...caused by failures in systems".[10(p146)] In other words, how we practice healthcare has a profound effect on the quality of care and consequently the outcomes of care. The purpose of this current study is aimed at exploring the concept of collaboration in healthcare and reporting on the validity of the dominant methods of measurement.

Population statistics indicate there is an increasing prevalence of chronic illness in modern societies. For example, the Australian Department of Health presents a summary of

data from the 2007-8 Health Survey on its website under the heading "Chronic diseases are leading causes of death and disability in Australia".[12] Despite improvements in the prevalence of cardiovascular disease (16% - down from 17% in 2001) and asthmatic conditions (10% - down from 12% in 2001) many indicators suggest that chronic diseases are on the increase; the prevalence of cancer (2% of the population – up from 1.6% in 2001), diabetes (4% - up from 2.9% in 2001), long-term mental or behavioural conditions (11% - up from fewer than 10% in 2001) and arthritis (15% - up from 14% in 2001). Mechanisms to improve quality of care are needed such as "...supporting integrated service provision and multidisciplinary care".[12] Due to its complex nature, chronic illness requires quality collaborative healthcare systems that effectively address this complexity. This further suggests that how we deliver healthcare influences the quality of care.

1.2 Caring and the healthcare setting

The act of caring is an important aspect of the process of curing illness.[13] It has been written that "...[c]aring for self and other human beings is a universal phenomenon that has endured beyond specific cultures, and has brought forth important humanistic attributes of care-givers and care-recipients...".[13(p57)] The environment in which this care is delivered is called the healthcare setting (HCS). For the purpose if this thesis, the HCS is any place where human health is the central activity of that setting and may include settings involved in the diagnosis and treatment of disease, the prevention of disease, the education of people to improve vitality and wellbeing, care of the elderly or disabled, palliation for people dying, and the rehabilitation of people with injury or post medical interventions. Therefore, the HCS may take innumerable configurations. However, common to all HCS is people who need healthcare (often called patients or care-receivers) and people who provide such care (carers or care-givers).

1.3 Identifying the attributes of HCS and practice styles

Validation studies (see Validation studies in Chapter 2) specifically evaluate

measurement properties of instruments which is the focus of this current study. Validation studies routinely report the characteristics of the HCS. Every HCS is unique and therefore the HCS should be considered a variable when research data is collected. It is posited here that the HCS can be defined by a set of attributes that are common to all HCS. This includes the participant types, the practice location and specialization and the style of team practice. Figure 1.1 represents the major attributes of the HCS which includes the location of practice (e.g. the country), the functions of the practice (e.g. surgery), the participants (e.g. professionals, non-professionals), the practice style (e.g. collaborative) and the care-receiver and family who, in this diagrammatic example, are placed at the centre of the HCS (representing patient-centred care principles).[14]

In this current study the attributes of the HCS are important in that they represent variables that place the results of individual validation studies within a defining context. The results of individual validation studies apply to that study. Therefore, generalizability of the study results to other HCS must be considered in reference to the attributes of the HCS (see Generalizability in Chapter 2). A validation study must report the study setting, ideally including a description of the location, functions, practice style and participants. In the systematic review reported in this thesis, only studies that adequately reported the HCS attributes were included. It is worth mentioning, that although practice style is put forward here as a HCS attribute, all studies included in the systematic review did not proffer a label for practice style, which might be valuable in future research.

Figure 1.1: Diagrammatized representation of the healthcare setting defining the nested characteristic variables of this generalized setting.

1.3.1 Participants - carers and patients

For the purpose of this thesis, carers may be professionals trained and qualified in various medical or therapeutic disciplines, some with highly specialized roles (e.g. anesthetist) and others with generalized skills (e.g. aged carer). Some carers may have no formal healthcare qualification (e.g. chaplain, orderly or patient's relative). It is also important to consider how each patient is unique relative to their age, gender, history, genetics, symptoms, culture and psychosocial characteristics. This is relevant if one intends to account for correlation attributable to patient variability. When measuring collaboration within HCS the effect of the patient's contribution to collaborative outcomes may be influenced by personal characteristics. For example, a patient who distrusts medical intervention may be antagonistic and unwilling to

collaborate, affecting collaborative outcomes.

1.3.2 Practice location and specialization

Throughout this thesis the reader will become aware of the breadth of HCS where the concept of collaboration needs to be effectively implemented and therefore effectively assessed. For example HCS are situated uniquely, relative to location (e.g. country, urban, rural) and specialization (e.g. surgery, general practice, rehabilitation, palliative care, natural medicine, community health, preventative medicine).

1.3.3 Team practice style

A healthcare team within a HCS does not always function collaboratively and team practice style varies between HCS. This variety can include parallel practice involving independent practitioners working in a common setting and performing their specific professional duties, developing their own treatment strategy without any input to other aspects of the patient's general healthcare. A patient may be visiting several practitioners simultaneously where each practitioner is conducting their own treatment plan for the patient's condition. Another practice alternative is called *integrative healthcare practice* (IHP).[15-21] These two examples of practice can be considered as two opposite practice styles on a continuum of collaborative team oriented healthcare. (see Figure 1.2).

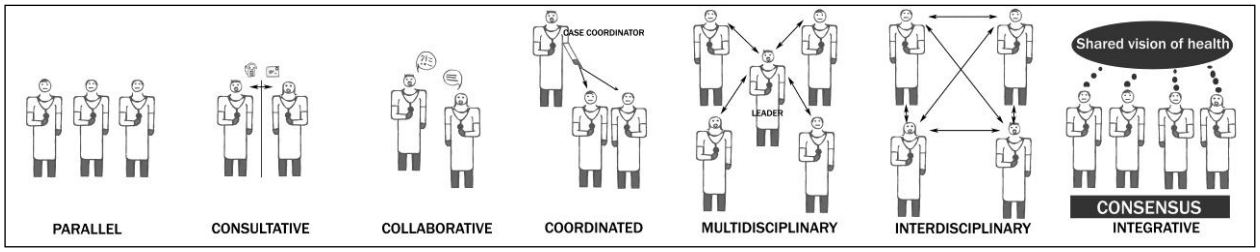

Figure 1.2: The continuum of team healthcare practices (adapted from Boon et al (2004)[21]

Boon et al.[21(p3)] defined IHP as;

> *"... an interdisciplinary, non-hierarchical blending of both conventional medicine and complementary and alternative health care that provides a seamless continuum of decision-making and patient-centred care and support... is based on a specific set of*

core values that include the goals of treating the whole person, assisting the innate healing properties of each person, and promoting health and wellness as well as the prevention of disease... employs an interdisciplinary team approach guided by consensus building, mutual respect, and a shared vision of health care that permits each practitioner and the patient to contribute their particular knowledge and skills within the context of a shared, synergistically charged plan of care".

Ideally, contributions to a patient's care should be collaborative. Collaboration is essential for integration. However collaboration may occur without integration. Research has shown that integrative care and collaborative care are perceived and defined by practitioners as two different but related social behaviors.[18] Where collaborative care is perceived as preserving practitioner autonomy while working with others, integrative care "...subsumes healthcare professionals under a common policy, organization, and structure".[18(p718)] Practitioners generally reflect an understanding that the goal of multidisciplinary practice is collaboration rather than integration.[18]

1.4 Definitions of terms used to describe teams

The terms multidisciplinary, interdisciplinary and interprofessional are commonly used throughout the healthcare literature to label the characteristics of healthcare teams. Such terminology however is not standardized and these terms are used with meanings relevant to an author's own interpretation.[22] However, for the purpose of this study the term *multidisciplinary* means the existence of carers working within a common healthcare setting who; have a unique discipline designation (referred to as a "participant type" in this current study) such as "physician", "nurse" or "social worker" for example; there are two or more disciplines within that healthcare setting; each carer is involved in the care of the same patient(s) and each carer works either in parallel or collectively.

Various definitions of the terms interdisciplinary and interprofessional exist in the literature.[22] For the purpose of this thesis the term *interdisciplinary* means where carers from multiple unique disciplines within the same professional group, such as medicine and may

include physician, nurse, anesthetist and oncologist for example, exist within the same healthcare setting and work collectively. For the purpose of this thesis the term *interprofessional* means where carers of multiple disciplines and from multiple professional groups work collectively. Such a mix may include for example community nurse, social worker, administrator and herbalist.

1.5 A definition of collaboration

There is a need to define collaboration, understand its constructs, the conditions under which collaboration occurs and standardize the measurement of collaboration. Development of an instrument to measure collaboration is difficult because of complexity and variability within HCS. The concept of collaboration has been developed by researchers and theorists across many disciplines. However, a coherent theoretical construct of collaboration should manifest regardless of discipline. In 1991 Donna Wood and Barbara Gray proposed a comprehensive theory of collaboration by examining inter-organizational collaboration.[2]

Wood and Gray extracted several definitions of collaboration from existing theory and synthesized a definition:

> *"Collaboration occurs when a group of autonomous stakeholders of a problem domain engage in an interactive process, using shared rules, norms and structures, to act or decide on issues related to that domain".* [2(p146)]

According to Wood and Gray[2] this definition encapsulates the essential elements of collaboration;

1. **Stakeholders:** of a problem domain; those interested in the problem. Do they all have the same interest or do their interests differ? Are all stakeholders involved or are some absent and what are the consequences of this absence?

2. **Autonomy:** even though stakeholders join the collaborative project they still maintain their decision making autonomy. They may surrender some of their autonomy but not all of it. If participants have no autonomous decision making capacity the structure is not collaborative;

3. **Interactive process:** collaboration is a process, a change oriented relationship in which all participants interact;

4. **Shared rules, norms and structure:** participants agree upon the rules and norms that will govern the interactive process. Structure is a shared structure and may be temporary, changing or permanent;

5. **Action and decision:** participants must have the intention to act and decide because a collaborative project has an objective to be reached. However, it is not essential for the objective to be reached for collaboration to have occurred;

6. **Domain orientation:** processes, actions and decisions must be directed towards the problem domain for which the collaborative project was established.

Wood and Gray emphasize that the outcomes of collaboration should not be integral in any definition of collaboration, stating "... a more general definition of collaboration thus should leave the consequences unspecified and open to empirical investigation".[2(p149)]

1.6 Measuring collaboration to improve teamwork

It is not sufficient to assume collaboration occurs without an empirical measurement. This current study aims to overview the measurement of collaboration in HCS by identifying and evaluating measurement instruments. By taking measurements, clinicians, managers and researchers can obtain a metric of teamwork that may translate to improvements in collaboration amongst the various team members, leading to increased patient safety,[23] better health outcomes[24] and enhancement of team member satisfaction.[25]

1.7 The theoretical basis for measuring collaboration in healthcare

Important to this current study is an understanding of the practicalities of quantifying phenomena that are not directly measurable. Collaboration is such a concept. Collaboration represents a complex social phenomenon that is not simple to quantify. For example, the collaborative Characteristic of Interdependence is one characteristic variable that cannot be

measured directly. The characteristic variables are called *latent variables* (alternatively called *factors*). Latent variables may be quantifiable by measuring observable phenomena (called *observable variables*) that have linear relationships with the latent variable. (For a discussion on latent and observable variables see the section on Validity). Therefore it is possible to indirectly measure Interdependence by directly measuring observable variables.

To quantify collaboration it is necessary to identify observable, empirically measurable variables. A theoretical understanding of collaboration assists in identifying the multi-faceted nature of collaboration and expands this concept into its component variables.

It is not an objective of this current study to elaborate on the theory of collaboration upon which the measurement of collaboration is based. Alternatively, the models used by the developers of measurement instruments will be briefly introduced here. These models are based on the extensive theory of collaboration discoverable in the literature.[2, 26] The dominant theoretical models upon which collaboration measurement in healthcare is based is defined in this current study by four approaches based on the publications of two independent researchers; Bronstein,[27] and Sullivan,[26] and two research teams; D'Amour et al.[28] and San Martín-Rodríguez et al.[29]

A note on terminology

Various terms are used in the literature to label the different variables identified or posited that relate to collaboration such as "characteristic",[26, 29] "element",[26, 29] "concept",[26, 28] "components",[27] "attribute",[26, 28] and "determinant"[29] which are adopted by authors depending on preference. For the purpose of this section, the term *concept* (from conceit)[30] is used as a global term for the abstract idea of collaboration as a phenomenon.[31] *Characteristic* is a distinctive essential trait[32] from which collaboration is comprised. An *Attribute* is a quality that forms or contributes to a Characteristic.[33] An *Element* is a "constituent portion of an immaterial whole, as of a concept or character[istic]".[31] It is proposed that Attributes and Elements are similar and for the purpose of this discussion the term Attribute will be used.

As an example of the use of this terminology, the collaborative Characteristic of Interdependence must exist for collaboration to occur. This Characteristic of collaboration has *communication* as one of its Attributes. Attributes are not unique to a particular Characteristic and so the Attribute *communication* for example, belongs to a number of different Characteristics. A Determinant is a general term that refers to any Characteristic or Attribute that influences collaboration in either a positive or negative way.[29]

To assist the reader in understanding the adopted nomenclature described above and presented within the theoretical models of collaboration discussed in the following sections, summary tables of the Characteristics, Attributes and Determinants of collaboration are given here (see Table 1.1 and Table 1.2)

Table 1.1 The Characteristics and Attributes of collaboration based on Bronstein,[27] Sullivan,[26] D'Amour[28] and Orchard et al.[34]

The Characteristics & Attributes of Collaboration	**Characteristics**	**Attributes**	**Author**
	Interdependence	*Communication*	Bronstein/Sullivan
		Mutual respect	Bronstein/Sullivan
		Constructed partnerships	Sullivan
	Newly Created Professional Activities	*Working together*	Bronstein
	Flexibility	*Role shifting*	Bronstein
	Collective Ownership of Goals	*Client centred care*	Bronstein
	Process/Reflection on Process	*Communication*	Bronstein
		Relationship	Bronstein
		Effectiveness	Bronstein
		Administrative support	Sullivan
		Leadership	Sullivan
		Resources	Sullivan
		Flexibility	Sullivan
		Commitment to collaboration	Sullivan
		Coordination	Orchard
	Partnership	*Sharing (see Organizational Determinants below)*	Sullivan/Orchard
		Respect	Sullivan/ D'Amour
		Communication	Sullivan/ D'Amour
		Working together	Sullivan
		Relationship	Sullivan
		Trust	Sullivan/ D'Amour
		Cooperation	Orchard
		Awareness of contributions of others	D'Amour
		Pursuit of common goals	D'Amour
	Power Sharing	*Competence*	Sullivan/ D'Amour
		Trust	Sullivan/ D'Amour

Table 1.2 The Characteristics and Attributes of the Determinants of collaboration - the presence/absence determines collaboration based on Bronstein,[27] Sullivan,[26] D'Amour et al.,[28] Orchard et al.[34] and San Martin-Rodriguez et al.[29]

	Characteristics	Attributes	Author
Characteristics and Attributes of Systematic Determinants	Power Disparity	*Gender stereotyping*	San Martin-Rodriguez
		Social status disparity	San Martin-Rodriguez
		Professional autonomy	San Martin-Rodriguez
	Professionalism/ Professionalization	*Domination*	San Martin-Rodriguez
		Autonomy	San Martin-Rodriguez
		Control	San Martin-Rodriguez
		Territorial behavior	San Martin-Rodriguez
		Conflict between professions	San Martin-Rodriguez
	Educational Induction	*Professional socialization*	San Martin-Rodriguez
Characteristics and Attributes of Organizational Determinants	Organizational Structure	*Heirarchial/horizontal (shared decision making)*	San Martin-Rodriguez
		Agency culture	Bronstein
		Administrative support	Bronstein
		Professional autonomy	Bronstein
		Time/space for collaboration to occur	Bronstein
	Organizational Philosophy (a positive belief in...)	*Participation*	San Martin-Rodriguez
		Fairness	San Martin-Rodriguez
		Freedom of expression	San Martin-Rodriguez
		Interdependence	San Martin-Rodriguez
		Openness	San Martin-Rodriguez
		Risk taking	San Martin-Rodriguez
		Integrity	San Martin-Rodriguez
		Trust	San Martin-Rodriguez
	Administrative Support	*Leadership*	San Martin-Rodriguez
	Team Resources	*Time*	San Martin-Rodriguez
		Space	San Martin-Rodriguez
		Financial	San Martin-Rodriguez
	Coordination and Communication Mechanisms	*Standards, policies and protocols*	San Martin-Rodriguez
		Unified/standardized documentation	San Martin-Rodriguez
		Sessions, forums and meetings	San Martin-Rodriguez
	History of collaboration	*Positive experience with*	Bronstein
	Sharing in partnership	*Shared work/intervention*	Sullivan/D'Amour

		Shared decision making	Sullivan/Orchard/D'Amour
		Shared problem solving	Sullivan
		Shared responsibility	Sullivan/D'Amour
		Shared goal setting	Sullivan
		Shared paradigm (vision, philosophy, values, professional perspective and ideas)	Sullivan/Bronstein/ San Martin-Rodriguez/D'Amour
		Shared planning	Sullivan/D'Amour
		Shared data	D'Amour
		Inclusion	D'Amour
Characteristics and Attributes of Interactional Determinants	Willingness to collaborate	*Group cohesion (constancy)*	San Martin-Rodriguez
		Open to the idea of collaboration	San Martin-Rodriguez
		Dedicated to the project	San Martin-Rodriguez
		Sharing	San Martin-Rodriguez
		Expressing a common objective	San Martin-Rodriguez
		Positive beliefs and expectations in collaboration	San Martin-Rodriguez
	Trust (in others & one's own ability)...	*Competence*	San Martin-Rodriguez
		Experience of the other professionals	San Martin-Rodriguez
		Confidence in one's own ability	San Martin-Rodriguez
	Communication	*Active listening*	San Martin-Rodriguez
		Effective and reliable systems of communication	San Martin-Rodriguez
	Interdependence	*Recognition and knowledge of the contributions of others*	San Martin-Rodriguez
		Mutual respect	San Martin-Rodriguez
	Professional roles	*Clarity of role expectations*	Bronstein
		Qualifications	Bronstein
		Values and ethics	Bronstein
		Allegiance to profession and Agency	Bronstein
		Respect	Bronstein
		Biopsychosocial view	Bronstein
		Shared vision	Bronstein
	Personal	*Understanding*	Bronstein
		Trust	

1.7.1 Bronstein Model of Interdisciplinary Collaboration

Laura Bronstein utilizes the definition of interdisciplinary collaboration by Bruner[35] as; "... an effective interpersonal process that facilitates the achievement of goals that cannot be reached when individuals act on their own".[27(p113)] Bronstein identifies the implied positive outcome of collaboration in this definition as opposed to collaboration being a neutral outcome similar to cooperation, communication, coordination or partnership. Bronstein's model draws on the literature in her identification of the central Characteristics of collaboration including Interdependence, Newly Created Professional Activities, Flexibility, Collective Ownership of Goals, and Reflection on Process.

According to Bronstein the Characteristic of Interdependence "...refers to the occurrence of and reliance on interactions among professionals where all are dependent on the others to accomplish their goals and tasks".[27(p114)] This Characteristic has the Attributes of *communication* between professionals; respecting of the opinions and skills of others (*mutual respect*) and the *understanding* of roles. Therefore, it is possible to conclude that by understanding Interdependence, all practitioners within the collaboration communicate openly with others, listen respectfully to the opinions and suggestions of others and possesses adequate knowledge of the paradigms, experience, methods and the role of each member of the collaboration.

The Characteristic of Newly Created Professional Activities results from the combined activities of all within the collaboration that amounts to a result that could not be achieved by the individuals alone (the whole being greater than its parts).[27] Collaboration facilitates the creation of new services that could not exist without collaboration. Therefore a collaborative healthcare team strives creatively to assemble their combined skill set to develop novel or customized approaches to a complex health problem.

The Characteristic of Flexibility "...is related to, but goes beyond interdependence to refer to the deliberate occurrence of role blurring".[27(p114)] Some Attributes of this Characteristic of collaboration include the capacity for *role shifting*, for example a therapist adopts the role of

case manager or a patient's family member is made responsible for managing the patient's wound dressing. Furthermore, Flexibility facilitates productive conflict resolution. Attributes such as *mutual respect*, *communication* and *clarity of role expectations* contributes to Flexibility and allows the collaboration to reconcile differences and resolve conflicts.

The Characteristic of Collective Ownership of Goals "...refers to shared responsibility in the entire process of reaching goals, including joint design, definition, development, and achievement of goals".[27(p114)] This Characteristic of collaboration refers to the Attribute *client centred care* in which the professionals, the patient and their families are all included in the process of goal attainment. Important to this Characteristic is each individual taking responsibility for their own role and "... behavior that supports constructive disagreement and deliberation among colleagues and clients".[27(p114)] Goals must result from a collective process of the entire collaboration. It goes to reason that goals set by only a subset of the collaboration may be unsatisfactory unless they are consensually acceptable to all members of the collaboration.

Bronstein's collaborative Characteristic of Reflection on Process involves "...attention to the process of working together".[27(p114)] Each member of the team evaluates their work critically so as to incorporate improvements and adapt to changing circumstances. Reflection on Process relies upon the Attribute of *communication* and it reflects the Attribute of *relationship* in collaboration.[27] Good *leadership* empowers the collaboration to act to make changes to processes. Leaders support the individual and facilitate change through upholding consensual decision making.

Bronstein identifies Determinants of collaboration that fall under the Characteristics of Professional Roles; Structural Characteristics; Personal Characteristics; and Positive History of Collaboration.[27p(115)]

Professional Roles has the Attributes of *clarity of role expectation*, *qualifications*, *values and ethics*, *allegiance to profession and agency*, *respect*; a *biopsychosocial view* and a *shared vision*, which defines

the Attributes needed by the healthcare professional to be an effective member of a collaborative team.

The Structural Characteristic has the Attributes of *agency culture*, *administrative support*, *professional autonomy*, and *time/space for collaboration to occur*. This Characteristic relates to the Organizational Determinant and represents the features an agency needs to facilitate collaboration.

The Personal Characteristic has the Attributes of *understanding* and *trust*. Each member of the collaboration must possess an understanding of the structure and function of the collaboration in all its aspects and display trust in the collaborative process.

The Characteristic of History of Collaboration has the Attribute of *positive experience with collaboration*. This Attribute suggests that a prior positive outcome of a collaborative project is an Organizational Determinant of collaboration. In other words, the experience of the organization or individuals within the organization has a potentiating effect on collaboration.

1.7.2 Sullivan's Critical Attributes of Collaboration

In Sullivan's model, collaboration is defined as "…a dynamic, transforming process of creating a power sharing partnership for pervasive application in health care practice, education, research, and organizational settings for the purposeful attention to needs and problems in order to achieve likely successful outcomes".[26(p6)] Sullivan performed concept analysis to reveal the Characteristics and Attributes of collaboration.[26(p6-42)] Concept analysis looks at the most common in-use concepts to understand any specific phenomenon.[26(p6-10)] As an alternative to creating inclusion/exclusion criteria for data, a consensual view of the current meaning of a concept is discovered.

The Characteristic of Partnerships has the Attributes of *respect* for each other, *communication, working together*, *relationship* and *trust* in each other.[26(p20)] Sullivan identifies the development of the Characteristic of Interdependence between collaborators as a result of *constructed partnerships* and as a product of *mutual respect* derived from *communication* behaviors.[26(p20-

[21)] Partnerships are dependent on *sharing* including *shared work*; *decision making*; *problem solving*; *responsibility*; *goal setting*; *vision, philosophy, values and ideas*; and *planning*.[26(p21)] The Characteristic of Process is a way of doing things as a progression. Central to this Characteristic are the Attributes of *administrative support*, *leadership*, *resources*, *flexibility* and a *commitment to collaboration*.[26(p14-15)]

Sullivan's research identified Power Sharing as a core Characteristic of collaboration strongly based on the Attributes of *competence* and *trust*.[26(p17)] "Each participant in a collaborative model must develop a high level of clinical knowledge and expertise in order to enter the relationship as a partner".[36(p28)] Confidence in the competence of the others is the foundation for participants to trust each other. Sullivan captures the essence of Power Sharing (and its antithesis - power disparity) by asking "...[s]hould the definition of collaboration be bold and embrace the term power sharing to best capture the idea of a partnership relationship in which two or more persons of differing but equally needed and respected abilities join together to address important human needs or problems?"[26(p17)] Underlying this question is the well researched issue of power disparity in healthcare practice. Research has shown that collaborative power facilitates a collaborative healthcare approach and is unlike the power exercised by medical doctors to dominate across various healthcare settings.[37]

Orchard et al.[34] developed a model for the development of the Assessment of Interprofessional Team Collaboration Scale (AITCS) based on the theory of Sullivan. Attributes of collaboration identified in this model include: *coordination* (the ability to work together to achieve mutual goals), *cooperation* (the ability to listen to and value the view based points of all team members and to contribute your own views), *shared decision making* (a process whereby all parties work together in exploring options and planning patients care in consultation with each other, patients and relevant family members), and Partnerships (creation of open and respectful relationships in which all members work equitably together to achieve shared outcomes).[34]

1.7.3 DÁmour's Conceptual Basis for Interprofessional Collaboration

Two constant and key outcomes in the formation of a collaboration were identified in the study conducted by D'Amour, Ferrada-Videla, San Martin Rodriguez and Beaulieu.[28] The first; collaboration is "...the construction of a collective action that addresses the complexity of client needs..." and secondly collaboration is "...the construction of a team life that integrates the perspectives of each professional and in which team members respect and trust each other".[28(p127)] These two principles are mutually dependent and provide some insight into the genesis of a collaboration, which is the development of a collective life to address the needs of the patient.

By reviewing the literature regarding the various definitions, concepts and theoretical frameworks of collaboration and via synthesis of the data, D'Amour et al.[28] proposed that collaboration is a complex construct comprised of five interdependent Characteristics; Sharing, Partnership, Power, Interdependency and Process.

According to D'Amour et al. Sharing is comprised of shared; *responsibility*, *decision making*, healthcare *philosophy*, *values*, *data*, *planning*, *intervention* and different professional *perspectives*.[28(p118)] Aspects of all these ways of sharing can be observed in a collaborative practice. For the purpose of this thesis and consistent with other authors, Sharing is an Organisational Determinant of collaboration. In other words, Sharing with its multiple Attributes can influence to what extent collaboration occurs. Sharing reflects the Attribute of *inclusion*; every member of the collaboration has the opportunity and responsibility to be involved with every aspect of the collaboration's processes. For example, in non-collaborative healthcare structures inclusion is prevented by hierarchical structures, where goals, planning and the prescription of interventions is controlled by experts such as a medical practitioner or a non-medical case manager. The Characteristic of Partnership "... implies that two or more actors join in a collaborative undertaking that is characterized by a collegial-like relationship that is authentic and constructive".[28(p118)] The collaborative Attributes of *communication*, *mutual trust* and *respect*, *awareness*

of the contributions of others and the *pursuit of common goals* and outcomes are associated with Partnership.[28] A partnership is not "authentic" if it exists in principle only. For example, a person is a member of the healthcare team but is not invited to contribute to case meetings - this is not Partnership.

"Collaboration requires that professionals be interdependent rather than autonomous and their interdependency arises from a common desire to address the patient's needs".[28(p119)] The word "autonomous" here refers to professional autonomy or independence and suggests a person acts alone and has no dependency on others, which is not true in collaboration. Some definitions of collaboration view autonomous decision making as a feature of collaboration[2] and therefore the difference between professional autonomy and autonomy in decision making is highlighted here. Synergy arising from the awareness of interdependency results in a maximization of individual's output (the whole being greater than its parts).

The Characteristic of Power Sharing amongst team members (power symmetry), is relational and based on each team member's knowledge and experience of each other.[28] Therefore power is intrinsically related to relationships within the collaboration. It is posited that quality leadership is a condition for collaboration because it goes some way in protecting power symmetry.[38, 39]

Each of the preceding models of collaboration help to emphasis the process/relational nature of team work that can be characterised as collaborative. Healthcare teams are collaborative when they are interactive, dynamic, transformational, interpersonal and able to transcend professional boundaries.[28(p119)] In other words, collaboration does not exist by name only. For example, calling a healthcare approach 'collaborative care' does not qualify the approach as collaborative unless the characteristics of collaboration are identifiable/quantifiable, again emphasising the need for a measure of collaboration.

1.7.4 San Martín-Rodríguez- The Determinants of Successful Collaboration

The Determinants of collaboration are those factors that influence the outcome of team

interaction and function. In other words collaboration is more likely to occur (or not occur) depending on the presence (or absence) or the quality of these Determinants. Determinants are described by their Characteristics and Attributes similar to the previous models described above.

By synthesizing empirical studies that evaluated collaboration, San Martin-Rodriguez, Beaulieu, D'Amour and Ferrada-Videla[29] identified three categories of Determinants contributing to successful collaboration. These were Systematic Determinants, Organizational Determinants and Interactional Determinants. Each of these categories is defined by multiple Characteristics and Attributes of collaboration (see Table 1.2).

1.7.4.1 Systematic determinants

San Martin-Rodriguez describes Systematic Determinants as those factors outside of the collaboration. The environment in which the collaboration exists is influenced by Systematic Determinants characterized by the social, cultural, professional and education systems which influence the outcome of collaboration.

The Characteristic of Power Disparity (power imbalance between professionals) includes the Attributes of *gender stereotyping* and *social status disparity* which are barriers to successful interprofessional collaboration.[29(p134)] Furthermore culturally based beliefs such as *professional autonomy* and those of different countries may also impede collaboration if those beliefs counter collaborative ideals.[29(p133-135)]

Professionalism also plays a role in impeding collaboration. In this context *professionalism* refers to "[t]he process of professionalization…characterized by the [Attributes of] achievement of [*domination*, *autonomy* and *control*], rather than collegiality and trust".[29(p136)] This is different to professionalism as behaving or performing duties in a professional manner. Professionalism may manifest as *territorial behaviors*, and *conflict between professions* due to ideological differences. For example, a physician may ignore or deride the opinion of a therapist based on a view that the medical profession is superior and therefore other opinions or treatments from

non-medical practitioners are inferior or flawed. Overcoming these professional barriers to collaboration may involve reflective practice (which fosters an understanding of the differences between professions) and developing a rationale for collaboration (rather than a rationale for professionalism) and social integration.[29(p137)]

Finally, Educational Induction is considered a major Characteristic of Systematic Determinants for interprofessional collaboration. Through the traditional processes of *professional socialization*, the opportunities for different professional groups to become familiar with the practices and ideologies of other professions was limited.[29] The emerging system of educating professionals to understand other professions is called Interprofessional Education (IPE) and is considered a way of nurturing professional pluralism (promotes sharing, awareness and integration of knowledge). This is an important Determinant in the development of collaboration.

1.7.4.2 Organizational Determinants

The organizational environment within which collaboration is to develop must be favourable. According to San-Martin-Rodriguez et al.[29] Organizational Determinants include the Characteristics of Organizational Structure, Philosophy, Administration, Resources and Coordination Mechanisms.

Organizational Structure needs to move away from the Attribute of a *hierarchical* to a more *horizontal* structure.[29(p138)] Practices that support teamwork and shared decision making are more likely to foster collaboration.[29(p139)] Hierarchy in healthcare is identified by the example of the legitimacy of a diagnosis in most medico-legal scenarios equating to that of the medical physician. The terminology equates 'medical evidence' to the doctor's opinion or diagnosis. Furthermore, treatment or diagnostic tests are only validated by the doctor's prescription or referral. In some healthcare settings, the capacity of non-medical personnel to dictate outcomes through control of funding for diagnostics and treatment is possible. If a case manager denies funding for treatment, perhaps on non-medical grounds, the prohibitive cost of such treatment

may prevent it. Hierarchical structure in healthcare settings facilitates power disparity and may be antithetical to autonomous decision making.

The Characteristic of Organizational Philosophy has the Attributes of "...*participation*, *fairness*, *freedom of expression* and *interdependence* ...a climate of *openness*, *risk-taking*, *integrity* and *trust* fosters collaborative attitudes between professionals".[29(p139)] Therefore an organization needs an explicit set of principles upon which collaborative practice is based and upon which the leadership of the collaboration functions.

The Characteristic of Administrative Support is essentially about the Attribute of *leadership*. Leaders convey the vision of collaborative practice and motivate people to adopt that vision. In other words, these leaders create an organizational setting that promotes collaboration. [29(p139)]

Adequate Team Resources are necessary for collaboration to develop. It has been shown that the Attributes of *time*, *space* and *financial* resources are important in this regard.[29(p139)] In other words, professionals collaborating require opportunities to share in order to address the problem domain collectively.

The Characteristic of Coordination and Communication Mechanisms is essential for successful collaborative practice. "Interprofessional collaboration can benefit, in particular, from the availability of *standards, policies, and* interprofessional *protocols*; *unified* and *standardized documentation*; and *sessions, forums and formal meetings* involving all team professionals".[29(p140)] This indicates that collaborative practice is highly organized; it has a well thought out structure, explicit processes and mandatory communication protocols.

1.7.4.3 Interactional determinants

The third category of Determinants within the San Martin-Rodriguez model[29] of collaboration is Interactional Determinants. This category includes the Characteristics of Willingness to Collaborate, Trust, Communication and the Attribute of *mutual respect*.

According to San-Martin-Rodriguez et al.[29] collaboration is, by its nature, voluntary.

Therefore, for collaboration to occur people must be willing to collaborate. Willingness to Collaborate has the Attribute of *group cohesion* (indicated by a constancy of professionals within the group) which is an important indicator of the willingness to collaborate. Professionals must be *open to the idea of collaboration* and *dedicated to the collaborative project*. *Sharing and expressing a common objective*, *positive beliefs and expectations* of the benefits of collaboration are also important.[29(p141)] Unwillingness to collaborate may threaten the success of a collaborative project. Willingness to collaborate may be cultivated by educating the group about the benefits of collaborative practice and creating opportunities to develop group cohesion.

For collaboration to occur, Trust is a key Characteristic of the Interactional Determinant and has the Attributes of *confidence in one's own abilities* and in *trusting others*. Trust in others is fostered by the Attributes of *competence* and *experience of the other professionals*.[29(p141)] Trust in others might be developed over time through the experience of working together but also through a good understanding of the experience, knowledge and skills others bring to the collaboration.

"Open and active communication and active listening... make[s] mutual knowledge possible among team professionals... and allows improvements to processes for sharing clinical information".[29(p142)] Without communication, it is unlikely a collaborative project can proceed successfully. As a Characteristic, Communication will depend upon the Attributes of *active listening*[40] and *effective and reliable systems of communication*, which may include face-to-face opportunities, electronic messaging and data sharing.

When working in a collaboration with others it is essential to have *mutual respect*, which "...implies knowledge and recognition of the complementarities of the contributions of the various professionals in the team and of their interdependence".[29(p142)] Having knowledge of what others contribute to the collaborative project and understanding how that contribution is important to the outcomes of the project is a Determinant of collaboration.

The theories upon which the measurement of collaboration is based are important in

the process of designing instruments as they guide the development of questionnaire and checklist items. This chapter has presented the theories that have a direct relationship to the measurement instruments reviewed in this study. To understand the theoretical construct of collaboration as a concept it has been helpful to further standardize the terminology used by various authors, hence the presentation in this chapter of a nomenclature based on Concept, Characteristics, Attributes and Determinants.

The next chapter provides the reader with a detailed discussion of the methodology and issues surrounding evidence synthesis. The chapter commences by defining evidence followed with a discussion of the JBI Model of evidence evolution and evidence generation. The central methodologies of evidence generation (qualitative, quantitative and validation research) are defined. The importance of evidence quality is discussed and the method of ranking quantitative evidence for therapeutic research is provided as an example. Evidence Based Healthcare is introduced before a discussion of evidence synthesis and the purpose of the systematic review in evidence synthesis. The following section introduces the principles of measurement science, validation studies, generalizability and the principle for the levels of evidence for instrument validity.

Chapter 2: Methodology

This chapter covers issues relating to the methodology used in this current study starting with the meaning of evidence and evidence generation. Following this, the methodologies for evidence synthesis are introduced followed by a discussion of the importance of evidence quality. Evidence based healthcare is introduced and the purpose of the systematic review in evidence synthesis is discussed. The principles of measurement science upon which this study relies is overviewed including test theories, validation studies, generalizability and the levels of evidence for instrument validity.

2.1 An overview of evidence in healthcare

2.1.1 What is evidence?

Evidence has been defined "…as a concept [that] pertains to truth, reality, and being in the world; it involves seeing, realizing, making visible, and clothing thoughts into words".[41] In a scientific sense evidence may be considered the testimony or facts supporting or disproving a hypothesis.[42] In other words evidence makes possible the realization of the truth value of a proposition. The term 'truth value' derives from Guba and Lincoln[43] who discuss the rationalistic approach to the nature of truth statements. They state that the "…aim of [rationalistic] inquiry is to develop a nomothetic body of knowledge; this knowledge is best encapsulated in generalizations (truth statements of enduring value that are context-free)".[43(p238)]

2.1.2 Evidence generation

Based on the JBI Model, evidence can be generated from research, discourse and experience.[5] The JBI Model proposes a four stage cycle of evidence evolution.[5] Relevant to this current study is the first stage; evidence generation. According to the JBI Model, evidence generation relates to evidence of Feasibility, Appropriateness, Meaningfulness and Effectiveness (FAME).[5] Feasibility relates to whether or not a healthcare practice is at all possible; appropriateness asks if a practice is suitable to the specific situation; meaningfulness relates to

the patient's positive experience of the practice; and effectiveness is a measure of the effect of the intervention related to health outcomes.[5]

It is widely acknowledged that evidence generated by research is the most reliable evidence in support of healthcare practice. Research may use various methodologies, for example qualitative, quantitative or validation.

In qualitative research the data collected relates more to the experiences or observations of the subjects or the researchers and are often generated as non-numeric or empirical forms of data. A definition of qualitative research is;

> *"…methodologies that provide holistic, in-depth accounts and attempt to reflect the complicated, contextual, interactive, and interpretive nature of our social world".*[44]

Quantitative research involves the collection of data using methods which involve taking measurements. Numerical data produced from measurement lends itself to analysis using statistical methods. Quantitative research has been defined as;

> *"… studies [that] produce results that can be used to describe or note numerical changes in measurable characteristics of a population of interest; generalize to other, similar situations; provide explanations of predictions; and explain causal relationships".*[45]

Validation research is concerned with the validity, reliability and interpretability of data produced by measurement instruments. In other words, is the data produced by an instrument measuring what it is intended to measure, is the measurement error within an acceptable range and can the data be interpreted in a meaningful way? Validation research is a type of quantitative research methodology because it uses numerical data in its analysis. However, quantization is not the only objective in validation research because some validation processes are more akin to qualitative methods. An example is content validation (see the section on measurement science later in this chapter).

2.1.3 The quality of evidence

Because evidence varies in quality it is necessary to rate the evidence so those who depend on it can make decisions or judgements with a degree of confidence. The quality of evidence relates to the quality of the methods used to generate that evidence. Consequently, ranking systems have been developed for a wide range of research methodologies. For example, when conducting research into therapeutic interventions it is widely considered that randomized controlled trials (RCTs) have a high degree of reliability due to reduction in bias. In contrast, the opinion of experts is less reliable. Therefore, it is possible to rank research and produce a level of evidence.

For example when ranking evidence generated from research of therapeutic interventions, a table published by the Centre for Evidence Based Medicine (CEBM - Oxford University) ranks the evidence on a five level scale.[46] The strongest level of evidence (Level 1A) is a meta-analysis of systematic reviews with homogeneous RCTs and the weakest level of evidence is expert opinion without explicit critical appraisal (Level 5). A diagrammatic representation of this hierarchy is presented in Figure 2.1

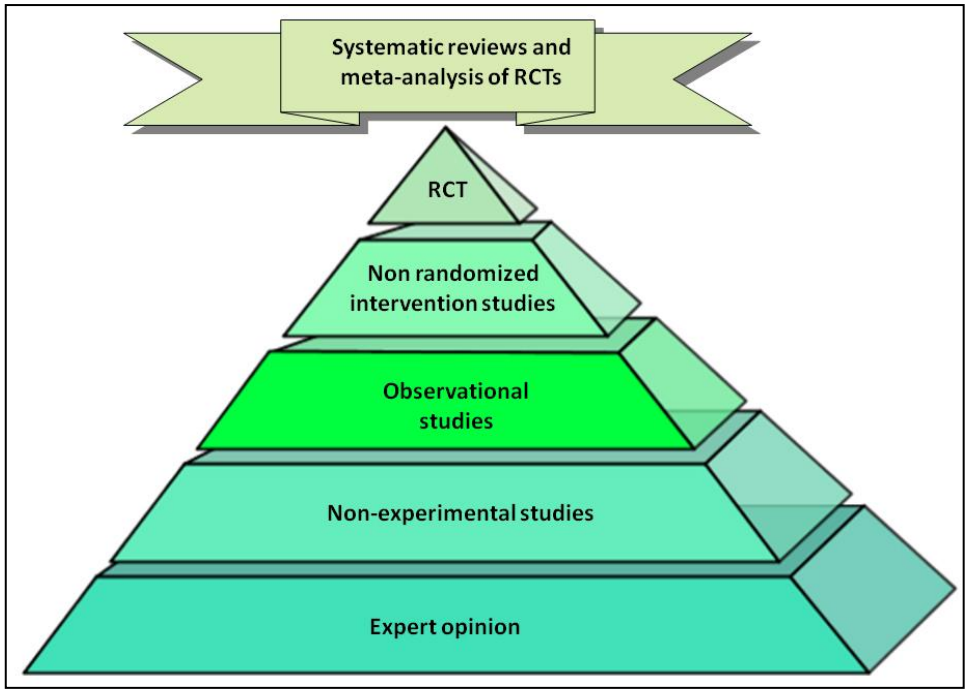

Figure 2.1: Hierarchy of study types for therapeutic interventions (adapted from Harbour and Miller 2001)[47(p335)]

2.1.4 Decision making in healthcare

The advent of evidence based medicine (EBM) has seen healthcare evolve from a science based predominately on principles of moral-ethical, legal liability, and economic rationalism[8] to decisions in healthcare informed by a careful consideration of the evidence. The Joanna Briggs Institute (JBI) Model conceptualizes Evidence Based Healthcare (EBHC) as "...clinical decision-making that considers the best available evidence; the context in which the care is delivered; client preference; and the professional judgement of the health professional".[5] Therefore evidence has become important in the process of decision making. This point is relevant to the current study because one pivotal source of evidence is measurement and the accuracy of measurement influences the quality of decisions that depend upon it.

2.2 Evidence synthesis

2.2.1 What is evidence synthesis?

In the JBI Model of EBHC, the second stage of the evidence cycle is evidence synthesis.[5(p211)] The product of the synthesized data may improve certainty when making decisions regarding a clinical or policy question in healthcare.[6]

Evidence synthesis is defined by Athanasiou and Darzi[6(p3)] as;

> *"...the synthesis (or integration) of variable data to produce information in the form of best evidence. It provides a set of methodologies to identify areas of agreement and disagreement in qualitative and quantitative data sets. By integrating data sets, this methodology may calculate the concordance and magnitude of effects from multiple studies".*

There are many ways to merge multiple studies in an evidence synthesis, for example systematic review, meta-analysis and narrative synthesis. The method used will depend on the source and type of data which may be produced by quantitative, qualitative or validation research.

2.2.2 Systematic review

2.2.2.1 The purpose of systematic reviews

The need for systematic reviews arises from the trend towards an evidence based approach to healthcare. Saso et al.[8(p71)] define a systematic review as;

> *"...the objective, transparent and unbiased location and critical appraisal of the complete scope of research in a given topic and the eventual impartial synthesis and, if possible, meta-analysis of individual study findings. Therefore, in order to address a specific research aim, a systematic review collates all evidence that fits pre-specified eligibility criteria".*

For example, the systematic review has advantages over RCT's in that data from multiple studies can be combined providing greater case numbers, increased statistical power, greater precision in the findings and hence greater certainty and less ambiguity regarding interpretation of the meaning of the results.[8] Increasingly, decision making is reliant upon the evidence that permits the best practice for the purpose of improving patient safety and health

outcomes. In order to use evidence for decision making with some confidence, it is essential that the evidence is presented in such as way as to be reliable. Relying on single study evidence is problematic due to issues relating to generalizabilty (see Generalizabilty later in this chapter). Furthermore, to gather evidence from multiple studies is not straightforward. This is due to the rigor and technical difficulty needed to search and integrate data stored in a vast and obscure network of literature, with conflicting findings and varying methodological quality. Saso et al.[8] encapsulated the problem as one of information overload, conflicting results, shortcomings of narrative review, limitations of RCTs and insufficient high quality trial data.

The systematic review has a defined process that reduces bias and appraises the methodological quality of the research to identify data that may be unreliable. The JBI Reviewer's Manual 2014[48] is one of many resources outlining the systematic review process. Where data from various sources can be pooled it may be possible to perform a meta-analysis which permits a summary about the effect size of an intervention compared to a control. However, a systematic review is not a meta-analysis. The result of a meta-analysis may be reported in a systematic review.

Meta-analysis is considered by Egger and Davis Smith[49] as a term used to describe the statistical integration of separate studies of quantitative data and for some time it was considered to be the systematic review. It is understood that meta-analysis is a "subset component" only of a systematic review.[8(p72)] With quantitative data produced by a series of RCTs it may be possible to estimate an average effect size using meta-analysis. However, it is essential to account for heterogeneity.

2.2.2.2 Heterogeneity

Higgins and Thompson[50] in their paper on quantification of heterogeneity in meta-analysis, stated "[a] systematic review of studies addressing a common question will inevitably bring together material with an element of diversity. Studies will differ in design and conduct as well as in participants, interventions, exposures or outcomes studied".[50(p1539)] This is known as

methodological or clinical heterogeneity which was encountered in this current study. Other types of heterogeneity include statistical heterogeneity, where the effects of an intervention differ between studies and where this difference is detectable if variation is greater than by chance only,[50] which was not a factor in the current study.

2.2.1.2 Narrative synthesis

Another type of evidence synthesis is called narrative synthesis and is "...considered to be 'typical systematic review', with clearly defined inclusion criteria and exclusion criteria in selecting and interpreting textual evidence...".[6(p13)] Narrative synthesis is essentially a summary of knowledge about a specified subject. Narrative synthesis was a methodology used in this current study because it was not possible to integrate the studies statistically. Data extracted from individual studies was integrated and presented as tables and figures and the characteristics and results of the studies were discussed. This data and the analysis were integrated and reported using a systematic review of measurement properties of instruments.

Therefore, a systematic review addresses these problems by making a synthesis of the research on a specific topic. By accessing systematic reviews, clinicians and researchers can make faster, more informed decisions regarding best practice.

2.4 Measurement science

2.4.1 Classical Test Theory

Measurement in science is based on several theories of measurement, two of which are important in the development of measurement instruments; Classical Test Theory (CTT) and Item Response Theory (IRT). Classical Test Theory evolved from the realization of the existence of error in measurement as a random variable and from the concept of correlation.[51] It is defined by the assumption that any measured score is the sum of the true score plus a measurement error and can be represented by the equation;

$$X = T + E$$

X = measured score

T = true score value

E = measurement error

The true score can never be known but it is assumed to exist within an interval. The measurement error is everything else measured minus the true score. Two categories of measurement error can occur; random (unsystematic) errors and systematic errors. Random errors are those created by factors that are inconsistent and may cause random deviations from the true score.[52] It can be assumed that over a large number of measurements, random errors, both negative and positive, will cancel and approach zero. Whereas systematic errors of measurement are errors that occur consistently and that are not measuring the trait the test is designed to measure.[52] It is also assumed in CTT that measurement errors do not correlate with the true score and therefore measurement errors in one test do not necessarily correlate with measurement errors of another test. The degree of measurement error reflects the level of reliability.

When measuring a large sample, the variance in scores obtained is comprised of the variance in the true score and the variance in measurement error;

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

Where σ_X^2 is the variance of the observed scores, σ_T^2 is the variance of the true scores and σ_E^2 is the variance of the measurement errors.[53]

A further assumption of CTT is described as *parallel test;* each item of a test is a test in itself of the latent variable. Therefore, the latent variable affects all items of a test in a similar way. It is further assumed that each item has the same amount of measurement error as other items of the test. In other words, all items of a test correlate to the true score. It is this

assumption that leads to the quantification of reliability. The parallel test model is considered to be a rather rigid model and other less rigid models exist. For example, the *tau-equivalent model* assumes each item has the same amount of error variance as the other items.[54] Another less rigid model is the *congeneric model* which assumes only that all items have a common latent variable.[54] Therefore, items may not have equally strong relationships to the latent variable and variances in error may not be equal.

2.4.2 Item Response theory

Item response theory (IRT) is a method of testing defined as;

> *"...a mental measurement theory based on the postulate that an individual's response to a test item is a probabilistic function of characteristics of the person and characteristics of the item".*[55]

The characteristics of the person relates to the person's trait level (latent trait), in other words their ability on the test.[56] The item characteristics relate to the difficulty of the item and its discriminating power. Item Response Theory testing developed significantly in the 1950's when computers could be employed to carry out the voluminous calculations required by IRT.[57]

Item Response Theory represents a diversity of different models,[58] all suited to differing applications depending on the characteristics of the item scoring, including the number of dimensions assumed to underlie performance, the number of item characteristics influencing responses, and the mathematical model relating the person's characteristics and the item's characteristics to the response.[55]

Some of the most common models include unidimensional models which assume that there is a single underlying trait to be measured. When the item is scored correct/incorrect the model is known as a *dichotomous model*. Multi-choice questions or a graded response fits a *polytomous model*. Other models include *continuous models* which are suitable where a measured response is used.[55]

Dichotomous models may incorporate one, two or three parameters. One parameter models assume the only parameter is the difficulty of the item. A two parameter model adds the

dimension of item discrimination which incorporates a capacity to discriminate between different individuals' ability. A three parameter model incorporates a parameter that allows a predication of the probability of guessing the correct answer for an individual with a very low level of the trait being measured.[55]

The equation representing a three parameter logistic model[59] estimating the probability of a correct response is;

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

i is the item

θ trait level (ability)

a_i, b_i c_i are the item parameters.

The curve produced by a IRT model is called an Item Characteristic Curve (ICC)[60] as illustrated in Figure 2.2 below. As the ability of the test subject gets greater, the probability of providing the correct answer gets also increases. The slope of the curve at P = 0.5 indicates the item's discriminating power. In this example Item 1 is easier to answer correctly than Item 2.

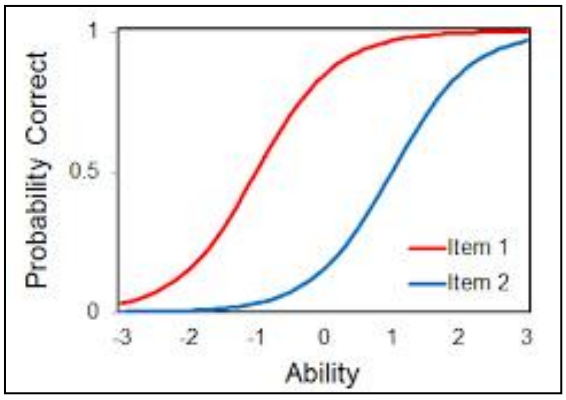

Figure 2.2: Item Characteristic Curve

Item Response Theory models are being used in increasingly diverse areas of

measurement including medical diagnosis.[61] For example, studies published in PubMed in 2014 indicate a growing interest in using IRT as a measurement model in healthcare.[62-83] In this current review, no instruments based on IRT were discovered.

2.4.3 Validity

In this current study the principal interest was the accuracy of measurement, specifically the validity and reliability of measurement instruments. Validity relates to the extent to which an instrument measures the construct it is intended to measure.[7] In the field of measurement instruments the concept of validity is comprised of sub categories; content and construct validity. Content validity addresses the question as to whether all aspects of an instrument are measuring the phenomenon of interest. For example, face validity tests the contents of the instrument by subjecting the instrument to the judgment of 'experts' in the domain in which the instrument is being utilized.

The experts may be clinicians, medical researchers or educators in their domain or they may be the target population of the instrument. Does the instrument, *on the face of it*, measure what is expected? Are any parts of it redundant, incomprehensible, irrelevant, or actually measuring something else but the phenomenon of interest? This qualitative approach helps test and refine the instrument to improve its content validity.

Besides using expert opinion only, content validity is often tested using a variety of methods. For example, Hull et al.[84] conducted a prospective, cross-sectional, observational study to test content validity of the Observational Teamwork Assessment for Surgery (OTAS) tool. Data collected from blinded parallel observations of two raters was subjected to statistical analysis to determine interrater agreement. Items on the instrument that showed poor interrater agreement were then subjected to expert scrutiny via a process of item ranking and subsequent refinement or removal of problematic items.

In another study testing the content validity of a culturally adapted version of the OTAS, Passauer-Bailer et al.[85] used semi-structured interviews with operating room experts and

used qualitative thematic content analysis to eliminate irrelevant items or misleading content and wording.

Validity also relates to the construct of the instrument. Do the scores of the instrument correlate with the stated hypothesis? Construct validity is categorized into structural validity, hypothesis testing and cross cultural validity.

In structural validity the question is; does the instrument scores reflect "...the dimensionality of the construct being measured?"[7(p743)] Structural validity may be tested by performing statistical analyses on data collected using the measurement instrument. A common approach to this validation is the use of latent variable methodologies.[86]

A problem arising in much measurement in healthcare is the impossibility of measuring some constructs directly. For example, it is not possible to measure collaboration via a unique parameter but it is possible to quantify this construct by quantifying the *latent variables*. Latent variables are variables that underlie a construct. Some of the latent variables associated with the construct of collaboration are communication, coordination and cooperation. Latent variables are not directly measurable but may be quantified by measuring *observable variables*. Observable variables are variables that are directly affected by the latent variable. For example, the statement "Patients/clients concerns are addressed effectively through regular team meetings and discussion" measures an observable variable reflecting the latent variable of communication in the Collaborative Practice Assessment Tool (CPAT).[87]

Observable variables are susceptible to measurement error and method variance (i.e. where different measurement methods produce different values).[88] The advantage of using latent variable methodologies is that it is possible to obtain a pure measure of a construct without measurement error or method variance.[86] Because measurement error is a type of unique variance, determining the shared variance of the observable variables makes it possible to produce a true measurement of the latent variable. There are two types of latent variable methodologies; exploratory and confirmatory. The value of these methodologies is in finding a

parsimonious solution for the latent variables, meaning when the shared variance of the observable variables can be attributed to the smallest number of latent variables.

In exploratory methods there are no *a priori* hypotheses about the number of latent variables or the associations between observable variables and the latent variables. The two most common methods are Exploratory Factor Analysis (EFA) and Principal Component Analysis (PCA). The essential differences between PCA and EFA are explained by Rencher and Christensen that in EFA the "...variables are expressed as linear combinations of factors, whereas principal components are linear functions of the variables...".[89(p475)] In PCA the attempt is to explain the total variance, not to explain the covariances as in EFA. No assumptions are required in PCA, principal components are unique whereas factors are subjected to an arbitrary rotation (in an effort to improve interpretability) and if the number of factors are changed the factors also change which is not the case in PCA.[89] Principal Components Analysis is considered the better method for exploring the underlying latent variable structure obtained from a large number of observable variables.[86]

Principal Components Analysis methods are technically complex and a detailed explanation is given by Abdi and Williams.[90] Essentially, PCA represents an analysis of both shared and unique variance between the observable variables. Principal Components Analysis is helpful when considering a large number of observable variables to discover a small number of groups of interrelated variables.[91]

Exploratory Factor Analysis is essentially a statistical technique that models the covariance of the observable variables using the factor loadings of the latent variables (factors), the residual variance and the factor correlations.[92] It is valuable because it can reduce a large number of observable variables to a smaller number of factors. For a concise description of the methods of EFA see Rencher and Christensen.[89]

In confirmatory methods such Confirmatory Factor Analysis (CFA) the latent variables and the observable variables associated with them are known. Furthermore, each observable

variable is associated with only one latent variable. Using a data set it is possible to perform a CFA to evaluate how well the acquired data fits the model. For a concise description of the method of CFA see Rencher and Christensen.[93]

An example of structural validity is the study conducted by Schroeder et al.[87] who performed both an EFA and CFA to validate the Collaborative Practice Assessment Tool (CPAT). The EFA was a part of an initial pilot study, the results of which aided in revision and modification of the CPAT. The second pilot using the revised instrument used CFA to validate the structure of the instrument.

With enough studies of acceptable heterogeneity, it is possible to perform a meta-analysis of factor analysis data.[94-98] However, it was not possible to do this in this current study due to methodological and clinical heterogeneity between studies. Studies varied in the characteristics of the sample, the location and type of healthcare practice and the instruments used for measurement.

To test construct validity using a hypothesis test, it is necessary to create *a priori* hypotheses about the strength and direction of the mean differences or correlations. Sample data is collected and used to test the hypotheses. By performing a prospective observational study, Sevdalis et al.[99] tested two hypothesis that stated 1) there would be stronger correlations between two experts scoring the OTAS checklist verses the correlation between expert and a novice rater and 2) expert rater scores would be significantly different compared to the difference between expert and novice raters. Data analysis of correlations between raters scores and analysis of differences in mean scores gave empirical support to the hypotheses and therefore provided evidence of the construct validity of the OTAS instrument.

If an instrument created for use in one population is to be used in another culturally different population it is important to test the validity of the instrument in a sample of the new population. This is known as cross-cultural validation. The process of translating a tool into a different language presents validity issues. It is not adequate just to translate the instrument

from the original language into the second language. Furthermore, the process must include multiple forward and backwards translations and expert assessment to authenticate an accurate translation, not just words, but also comprehension of core constructs and their cultural correctness.[100, 101] Therefore content validation is important in the translated version of the instrument. Furthermore, the translated instrument needs to be tested for functional and measurement equivalence. In measurement equivalence (alternatively measurement invariance) the question is when, under different conditions such as different cultures, is the instrument measuring the same constructs in both situations?[102]

In a systematic review of the measurement properties of instruments, it is important to report the characteristics of the studies included in the review as this addresses the property of generalizability (see section Generalizability later in this Chapter) of the instruments. The characteristics of the study such as the participants (sample size, gender, age, and educational qualifications etc.), location of the study and cultural factors all influence the validity of the data produced. Validity data produced by a single study is unique to the sample and setting used in that study. In other words, the evidence reported by any single validation study is applicable to that sample only. When the instrument is applied to a different sample the validity and reliability of the instrument and hence the capacity to interpret the results of measurements is unknown. It is justifiable to assert that validity and reliability tests should be performed whenever clinical or research data is produced.[103]

2.4.4 Reliability

Any instrument must be reliable to be valid, but reliability does not depend on the instrument's validity. Reliability addresses the issue of consistency and therefore measurement error. As discussed previously, Classical Test Theory is based on the principle of observed scores being the true score +/- a measurement error. Reliability asks the question; if the phenomenon of interest does not change how consistent are measurements taken on different occasions. For example, how similar are the results if two observers rate an event

simultaneously (interrater agreement) or if a single observer rates an event on different occasions (intrarater agreement) or over a period of time (test-retest)?

A sub-type of reliability is internal consistency; a measure of interrelatedness between the scores of items of an instrument.[7] If the instrument is measuring the same construct (unidimensionality), high levels of correlation between the scores of items is expected. The most commonly used measure of interrelatedness is Cronbach's alpha.[104]

> *"... alpha is ...an estimate of the correlation between two random samples of items from a universe of items like those in the test".*[104(p297)]

The formula for alpha is for items(i) 1,2... n;

$$\alpha = \frac{n}{n-1}\left(1 - \sum_{i}\frac{Vi}{Vt}\right)$$

n = number of items

Vt = variance of the test scores

Vi = variance of the item scores after weighting

Reliability is an estimate of measurement error; squaring the reliability coefficient and subtracting from 1.0 produces an index of the measurement error.[105] Therefore the higher the reliability of the instrument the lower is the measurement error.

Other classes of reliability use different statistical methods for analysis. For example, interrater (alternatively interobserver) agreement may be reported as a percentage agreement (the percentage occurrences of agreement). However, when two raters make a judgment of the same phenomenon the possibly of agreement or disagreement by chance alone needs to be quantified. This can be tested using Cohen's kappa. Kappa is intended to give a quantitative measure of the magnitude of agreement between observers.[106] The formula for kappa is;

$$k = \frac{Po - Pe}{1 - Pe}$$

Po = proportion of the observed agreement between two raters

Pe = proportion of rater agreement by chance alone

Interrater reliability is also dependent on the consistency of scoring. According to Multon, "...consistency estimates of interrater reliability are based on the assumption that it is not necessary for the judges to share a common interpretation of the rating scale, as long as each rater is consistent in assigning a score to the phenomenon".[107(p5)] The three types of consistency estimations are correlation coefficients (Pearson's *r* or Spearman's rho[38]), Cronbach's alpha[104] and intraclass correlations (ICC).

Pearson's *r* is used to calculate reliability for one pair of raters and for each item at a time. This assumes a normal distribution. Where there is no normal distribution, Spearman's rank coefficient is used. Where more than two raters are used Cronbach's alpha should be used. Interval and ordinal data is best analyzed using ICCs which are very useful where variables of the same class share a common metric and variance (i.e. measuring the same thing).[108]

For example, the OTAS checklist reliability is strongly dependent on the ability of two blinded raters to rate the same observed procedure in the operating room and achieve approximately the same scores. Statistical correlation tests such as intraclass correlations, Cohen's kappa, Pearson's *r* and Spearman's rho were used to estimate interrater reliability of the OTAS instrument.[84, 99, 109-111]

2.4.5 Interpretability

Interpretability is not a measurement property in itself, however it is an important feature of an instrument. Interpretability is defined as;

> *"...the degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations – to an instrument's quantitative scores or change in scores".*[7(p743)]

Interpretability of an instrument relates to the usefulness of the instrument. Data that has been collected and analysed must translate meaning if it is to be useful. If it is not possible

to make a judgement from the data collected by an instrument, the instrument has poor interpretability.

Important to interpretability are the mean and standard deviation of the data and a description of the distribution of data (e.g. normally distributed). In this current study, research was included that reported the results such as means and standard deviations of assessments and interpreted the differences in the means to derive a qualitative understanding of the data. This research contributes to the interpretability of the instrument.

2.4.6 Validation studies

In this current review the focus is the accuracy of measurement, more specifically does the instrument measure what it is expected to measure (validity) and are the measurements reproducible (reliability)? Studies reporting on validity and reliability of instruments are called validation studies. Validation studies of measurement instruments are unique in the scheme of quantitative research evidence. In most quantitative research it is an objective to observe or compare some effect, like the therapeutic benefit of a drug verses a comparator treatment or the result of implementing a health policy on disease reduction. In validation studies, it is the accuracy of measurement that is under critique. One pressing issue with single validation studies is the limitation on the range of measurement properties assessed in any one study. A measurement property of an instrument is a specific indicator of the instrument's capacity to measure a variable with a determined level of accuracy within a specified context or setting. Another important limitation of single validation studies is generalizability.

2.4.6.1 Generalizability

When an investigator wishes to generalize the results of a study to a wider population of people, the issue of *validity* is relevant.[112] According to Tebes,[113] Cook and Campbell[114] classified four types of validity: internal validity; statistical validity; external validity; and construct validity. External validity is about generalizing inferences "to other persons, settings or times". [113] In this current study, generalizability relates to external validity. Specifically, the question arises; can a

measurement instrument validated with one sample of the population be used with another sample and produce measurement results that can be relied upon to make a judgement?

Generalizability issues arise from the fact that the results of any single validation study relate to the sample used for that study only. Furthermore, most single validation studies address specific measurement properties and do not address other properties. For example, in the initial stages of developing an instrument to measure specific psychometrics properties the concepts of internal consistency, content validity and structural validity are priorities. However, the measurement properties of construct validity using hypothesis testing, criterion validity, test-retest/inter-rater reliability, cultural validity and responsiveness are also required. Resources to assess each level of validation may not be available for a single study.

2.1.6.2 Level of evidence for instrument validity

The level of evidence of the validity and reliability of a measurement instrument is partly determined by the process of a validation study, which is directly related to the methodological quality of the study. Furthermore, the level of evidence also relates to the consistent reproducibility of validity data, meaning that multiple studies producing consistent results between studies increase the level of evidence. By performing a systematic review of measurement properties of instruments it is possible to produce a synthesis of multiple validation studies and determine the level of evidence for validity and reliability of measurement instruments.

This chapter has presented the methodological principles relevant to the current study. The principles upon which evidence synthesis, systematic review and measurement science is based were presented. These principles underpin validation research and the systematic review of measurement properties of instruments.

In the next chapter, the method used to undertake the systematic review of measurement properties of instruments is detailed. This includes an overview of the systematic review method, assessment of methodological quality using the JBI approach, the use of the

COnsensus on the **S**election of **M**easurement **IN**struments (COSMIN) checklist to assess the methodological quality of validation studies, the method of evidence synthesis of validation studies, and assigning a level of evidence to this type of review. This is followed by a facsimile of the published protocol for the systematic review that details the criteria of the study including the background to the study, the initial search strategy, inclusion/exclusion criteria, assessment of methodological quality, data collection and synthesis.

Chapter 3: Methods

This chapter details the methods involved in conducting a systematic review of measurement properties of instruments. It commences with an overview of the systematic review method and the systematic review protocol. Following this is a description of the systematic review of measurement properties, the method of assessing methodological quality using the COSMIN checklist, the method of evidence synthesis of validation studies and ascribing a level of evidence. Finally, the published review protocol used to guide the review process is presented.

3.1 The systematic review

3.1.1 Method

The method of conducting and reporting a systematic review is described in the JBI Reviewers Manual 2014.[48] The steps involved in the production of the current systematic review were;

1. Publication of a systematic review protocol that outlined the essential objectives and methods of the review.
2. Performed searches using the search terms stated in the protocol within the targeted databases and journals.
3. Merged results of searches and removed duplicates and studies that did not match the inclusion criteria.
4. Systematically evaluated all included studies for methodological quality (critical appraisal) using standardized and validated appraisal tool(s). This involved rating each study for methodological quality based on predetermined parameters that reflected the research objectives.
5. Extracted data from all included studies based on the predetermined requirements.

6. Performed a data synthesis based on the appropriate method dependent on study or data heterogeneity. This could have been a meta-analysis or narrative synthesis for example.

3.1.2 The systematic review protocol

The guidelines for a systematic review are set out in The JBI Reviewer's Manual[115] and were adopted in the process of completing the systematic review presented in this current study (see Chapter 4). The subsequent systematic review process followed the guidelines of the protocol.

The systematic review protocol was a separate publication to the systematic review report. The protocol was important as it provided a transparent process upon which the systematic review was structured which helped reduce reporting bias. The systematic review protocol set out the reasons for and the method of performing the systematic review. For this current study the JBI guidelines for a systematic review protocol[115(p52-65)] were followed except where specific processes relating to the systematic review of measurement properties of instruments required adoption of tools not provided by the standard JBI tools.

The initial step in this process was the development and publication of a systematic review protocol. The stages involved in this process were;

1. Provide a rationale for the research with reference to the published research (background).
2. Define an inclusion/exclusion criteria for studies (**P**opulation, phenomenon of **I**nterest, **C**omparator , **O**utcome of interest - PICO).
3. Define search criteria and identify relevant databases.
4. Adopt the appropriate appraisal tools for assessing methodological quality.
5. Expert and peer review, protocol refinement and publication.

3.1.3 The study report

The systematic review report should consist of essential sections; the title, executive summary, background to the study, inclusion criteria, search strategy, method (methodology, data collection, data synthesis), results (description of studies, PRISMA diagram representing the results of database searches and the process of inclusion and exclusion of studies, appraisal of methodological quality of included studies, results of the data synthesis), discussion and conclusion (implications for practice and research) and references. Supplementary data is attached to the report including tables of excluded studies, the characteristics of included studies and appraisal and data extraction tools.

The JBI has strength in the publication of systematic reviews in healthcare and provides software tools for the management and publication of systematic reviews. The JBI **S**ystem for the **U**nified **M**anagement, **A**ssessment and **R**eview of **I**nformation (SUMARI) assists researchers and practitioners to appraise and synthesize evidence in health and social science. However, for this current study the tools required for appraisal and synthesis were not available in SUMARI.

3.2 Systematic review of measurement properties

A systematic review of measurement properties is a review of all available studies on the measurement properties of all available instruments that aim to measure a particular construct in a particular population.[116] This type of review can take various configurations for example; all studies available that measure a specific construct for one instrument, or a selection of the most commonly used instruments, or all instruments that measure a particular construct, or all instruments for a particular population without specifying the construct. Developers of instruments report on the developmental process in relation to the theory, sample description, construction, validation and interpretation of the tool. A systematic review of measurement properties of instruments searches for these reports, extracts data relevant to these variables and

presents a synthesis of the results. In the current study the approach used was to evaluate all validated instruments that measured collaboration in HCS with a complex mix of participant types.

A tool for the critical appraisal of studies about measurement properties was not currently available in JBI SUMARI. Therefore, a decision was made to use the **CO**nsensus on the **S**election of **M**easurement **IN**struments (COSMIN) checklist for the appraisal of the methodological quality of studies reporting measurement properties and for data extraction relating to study participants and interpretability.[7, 117, 118]

3.2.1 Consensus on the Selection of Measurement Instruments checklist

In 2007 Terwee et al.[119] proposed that similar to systematic reviews of clinical trials, when comparing measurement instruments of health outcomes, studies must be rated for methodological quality. A rating system based on the appraisal of methodological quality was suggested.

The COSMIN checklist is a standards assessment tool and was developed by a panel of international experts using a Delphi study in 2006.[117] The aim of this initiative was to reach consensus on which measurement properties should be evaluated and the definitions of those properties, when examining health related patient-reported outcomes (HR-PROs).[7] Furthermore, how these measurement properties should be evaluated relative to study design and statistical analysis.[120] A further development of the COSMIN checklist allows reviewers to rate the methodological properties of a study using a four point scale.[121]

The COSMIN checklist consists of twelve boxes (see Appendix 1). Ten of these boxes are used to assess the methodological quality of a study. Nine deal with measurement properties; internal consistency (Box A), reliability (Box B), measurement error (Box C), content validity including face validity (Box D), construct validity i.e. structural validity (Box E), hypothesis testing (Box F), cross cultural validity (Box G), criterion validity (Box H) and responsiveness (Box I). Box J contains standards for studies on interpretability. In addition to

these ten boxes are two boxes, one assesses articles using Item Response Theory (IRT box), and the second box is for determining the generalizability of results of a study to one or more measurement properties (Generalizability box).

The COSMIN checklist is designed as a modular tool. This is a useful feature as individual studies may not assess all measurement properties. For example, a study may assess internal consistency and reliability but not cross cultural validity. Therefore Box A and Box B would be completed but not Box G. Multiple groups or subgroups may be included in some studies, so the same box may be completed multiple times. A modified form of the generalizability box was used in this study to compile characteristics for each study sample (see Appendix 2).

In assessing methodological quality, each study is appraised by two independent appraisers. Results of appraisal are compared and any discrepancies being discussed and resolved between the two appraisers. If this is unsuccessful, a third appraiser is used to resolve the discrepancy.

3.2.2 Evidence synthesis of validation studies and reporting

The JBI 2014 Reviewers' Manual[115] provides guidelines for the preparation of reports of systematic reviews. The manual outlines the layout of a review report of quantitative data around which this report was structured. Due to the specific nature of this current type of review some additional conventions were adopted.

Since the introduction of the COSMIN checklist there have been a number of recent publications of systematic reviews that have used the COSMIN tool.[122-125] A precedent for reporting this type of review has consequently been established. This current review has adopted report features based on Paalman et al.[124] There is currently no review in the JBI Library using this type of review methodology. This adds to the current JBI report format. This includes a best evidence synthesis based on the criteria by Terwee et al.[119] who proposed; "in the final comparison of the measurement properties of different questionnaires, one has to

consider all ratings together when choosing between different questionnaires. We recommend to compose a table that provides an overview of all ratings...".[119(p38)]

Statistical pooling of data was not possible for this review. Therefore, the findings are presented in a narrative form including tables and figures to aid in data presentation. To make a synthesis of the assessments of methodological quality of the different studies, we rated each instrument using the scheme proposed by Terwee et al.[119] and utilized by Paalman et al.[124] By accounting for the number of studies performed with an instrument, the appraisal of methodological quality and the consistency of the results between studies it was possible to rate the instrument. This is similar to the Cochrane Collaboration Back Review Group guidelines.[126, 127]

3.2.3 Levels of evidence of validation studies

Ascribing the strength of evidence to support validity (levels of evidence) for this type of review is unique. Some recent studies have utilized a scale for rating the levels of evidence and this scale was adopted for this current review (see Table 4.2).[122, 124, 128] The precise method of rating has been faithfully adopted from the study by Paalman et al.[124]

This scale utilizes a coding system to record the result of the appraisal of methodological quality for each measurement property assessed plus accounts for the consistency across a number of studies performed on the same instrument. The coding system utilizes three symbols; +/-, ? and NA. Where evidence is provided within a study addressing instrument validity, and depending on the strength of that evidence, the symbol '+' is tabled indicating positive support of validity. The symbol '-' would indicate evidence but of negative support for validity. If evidence for validity was weak and therefore unknown, the symbol '?' would be tabled. 'NA' indicates the measurement property was not assessed.

3.3 The measurement of collaboration within healthcare settings: a systematic review protocol of measurement properties of instruments

This section reproduces the systematic review protocol published in the The JBI Database of Systematic Reviews and Implementation Reports.[129]

3.3.1 Review Question(s)/Objective(s)

The objective of this review is to evaluate and compare the measurement properties of instruments that measure collaboration within healthcare settings, specifically those which have been psychometrically tested and validated.

More specifically, the objectives are to:

1. Identify studies reporting the measurement properties of instruments that measure collaboration within healthcare settings that are populated with a complex mix of participant types.
2. Identify the measurement properties assessed by each study.
3. Evaluate the reports on methodological quality and rate them.
4. Compare instruments by synthesizing the results of the evaluation.

3.3.2 Backgound

It has been stated that the idea of teamwork and collaboration in the healthcare setting (HCS) is intuitively appealing.[130] However, research and general experience indicates that the achievement of teamwork and collaboration is modest in the majority of HCSs[131] with the perception and experience of collaboration often varying between professionals working in the same setting.[132]

The term *team* is difficult to define as a universal entity. In the literature several terms are used to label types of teams within HCS such as multidisciplinary, interdisciplinary and inter-professional.[22] These terms commonly target the health professional groups within the

HCS and are not inclusive of the patients themselves, their friends and family or other types of non-professional groups involved in the care of the patient. For this reason we will focus on the *participants* within HCSs and not exclusively on inter-professional teams. Any real HCS is likely to be populated with various types of participants including orderlies, receptionists, chaplains, clerical staff, administrators and volunteers who may all contribute to a patient's care. The impetus to consider others in the HCS such as the patient and their families redefines the carer team.

Concepts like shared decision making;[133] involving patients in safe care approaches to inter-professional practice;[134] patient and family involvement in quality improvement processes[135] and the World Health Organization's[136] call for patient and family inclusion in collaborative healthcare all reflect a growing awareness of the need to understand and collaborate with others within the HCS. Therefore, collaboration in the HCS is best considered to be broader than the 'professional' groups (e.g. nurses, physicians and pharmacists etc.).

A review of the existing research and discourse on collaborative teamwork in healthcare suggests that the presence of collaboration can result in improving patient outcomes and enhancing team members overall levels of satisfaction.[24, 25] For example, patient safety in relation to drug prescription improves when nurses and pharmacists collaborate.[23] Routinely, different professional groups work in teams for example in surgery where the surgeons, anesthetist and nurses etc. work as a team to achieve specific goals. However, can this teamwork be considered collaborative?

The term/concept 'collaboration' is often used in the literature and adopts various meanings depending on the author's viewpoint and the context or environment in which the team operates. Barbara Gray[137] defined collaboration as the process of joint decision making by interdependent stakeholders involved in solving a specific problem. Gray suggested that collaborative decision making involves stakeholders resolving differences, joint ownership of the decisions reached and collective responsibility. In an editorial published in 2000 titled

'What's so great about collaboration? We need more evidence and less rhetoric', Zwarenstein and Reeves[138] highlighted the need for more research to justify the application of collaboration in inter-professional healthcare practice. The interest of this current review is the tools used to measure collaboration.

A current search of the literature indicates a significant research effort into the outcomes of collaborative healthcare. A deficiency in the collaborative care research is to associate positive patient outcomes as a result of collaborative care.[139-144] However, without an empirical measurement the observed outcome may be due to a multiplicity of variables. In a Cochrane systematic review, Zwarenstein, Goldman and Reeves[24] identified five randomized controlled trials of Inter-Professional Collaboration (IPC) interventions and concluded IPC was effective in improving healthcare outcomes. Only one study cited in the review attempted to evaluate team collaboration by comparing the measured outcomes of videoconferencing and audio-conferencing.[145] The review authors stated "... we know little about the processes of collaboration and how it contributes to changes in healthcare processes and patient outcomes".[24(p8)] The authors suggest that there is a need for "...future research... [to] ...focus on the conceptualizations and [validation of] measurement [criteria] of collaboration".[24(p9)]

A number of theoretical models of collaboration have evolved within the broader framework of human behavior that assist in understanding the group behaviour of collaboration.[146] Relevant to the healthcare and social care settings are three theoretical models that attempt to define and conceptualize collaboration; Sullivan, [26] D'Amour[146] and Bronstein.[27] Theorization and conceptualization assists in the identification of the key determinants of successful collaboration[29] and in turn, the measurement of collaboration.

According to Orchard et al.[34] Sullivan's model is based on the "...critical attributes of collaboration..." *coordination* (includes achieving mutual goals by working together), *cooperation* (contribution of views and valuing those of other team members), *shared decision making* (planning care in consultation with all including the patient and their families) and partnership

(creating effective working relationships).

D'Amour's model[146] is based on the outcome of a synthesis of 17 papers regarding collaboration. The attributes of collaboration identified in this model include *sharing* (responsibility, decision making, healthcare philosophy, values, data, planning and interventions), *partnership* (collegial relationship that involves open communication, mutual respect and trust; value the contribution of others and common goals), *interdependency* (mutual dependence = the whole is greater than the sum of its parts) and *power* (symmetry in power relationships).

Bronstein's model[27] includes the collaborative attributes of *interdependence*, *newly created professional activities* (new activity and services not achieved without collaboration), *flexibility* (the deliberate occurrence of role blurring), *collective ownership of goals* (shared responsibility in the process of reaching goals) and *reflection on process* (attention to the process of working together).

In addition to models and attributes of collaboration the factors that promote or impede collaboration need to be considered when attempting to measure collaboration. A 2005 review of the literature resulted in the identification of three determinants of successful collaboration; systematic determinants, organizational determinants and interactional determinants.[29] Each of these determinants is dependent on a multiplicity of factors. For example, the systematic determinant is influenced by the social, cultural, professional and education systems. The organizational determinant is impacted by the organization's structure, philosophy, administration, resources and coordination mechanisms and the interactional determinant is influenced by peoples' willingness to collaborate, trust, communicate and mutual respect.[29]

Research into healthcare team collaboration has relied upon the adaptation of existing instruments to measure collaboration. These instruments are not specific to inter-professional teams and few have been validated psychometrically. Orchard et al. suggest that instruments which allow "...teams to assess collaborative relationships are needed".[34(p59)] Thannhauser,

Russell-Mayhew and Scott[147] evaluated twenty three instruments measuring inter-professional education and collaboration. This evaluation included development of psychometric properties, validity and reliability data, general utility of the measure, sample description and questionnaire design which are also important criteria for this review.

Instruments such as the Index of Interdisciplinary Collaboration (IIC)[27] and its modified formats (Modified Index of Interdisciplinary Collaboration-MIIC) have demonstrated a capacity to measure and differentiate variances in the perception of collaboration within a hospice setting[148-151] and measure collaboration in expanded school mental health programs.[152] Other instruments such as the Inter-professional Socialization and Valuing Scale,[153] the Assessment of Inter-professional Team Collaboration Scale (AITCS),[34] the Care Process Self-Evaluation Tool (CPSET),[154] the Doctor's Opinion on Collaboration (DOC)[155] and others also exist, however no systematic reviews have been conducted to evaluate these tools.

For the purpose of improving patient safety, improved collaboration between people within any HCS needs to be facilitated. For example, Dougherty and Larsen[156] reviewed measurement instruments for nurse-physician collaboration and recommended collaboration as a key communication strategy to minimize errors and increase patient safety. Healthcare policy makers and administrators are increasingly promoting collaborative teamwork as a key foundation of effective and efficient healthcare. Given the acclaimed role that collaboration plays in improving patient safety and health outcomes, it is important to determine effective ways to measure collaboration in the HCS. Research outcomes are invalid if there is an assumption that collaboration has occurred without an associated measurement using a validated instrument. The purpose of this review is to identify which of the available instruments are valid and reliable measurements of collaboration in the HCS populated by a complex mix of participant types.

3.3.3 Criteria for considering studies for this review

3.3.3.1 Types of studies

The types of studies considered for inclusion will be validation studies, but quantitative study designs such as randomized controlled trials, controlled trials and case studies are also eligible for inclusion. Studies that are Interprofessional Education (IPE) focused, published as an abstract only, patient self-reporting only or not about care delivery are also excluded.

3.3.3.2 Types of participants

Participants may be any healthcare professionals, the patient or any other non-professional who contributes to a patient's care. The term *participant type* means the designation of any one participant; for example 'nurse', 'social worker' or 'administrator'. More than two participant types is mandatory. Diversity of participant types includes the diversity observed between medical doctors, for example oncologist, radiologist or general practitioner.

3.3.3.3 Focus of the review

The focus of this review will be the validity and reliability of instruments used to measure collaboration within healthcare settings.

3.3.3.4 Types of outcome measures

The outcome of interest is validation and interpretability of the instrument being assessed that includes content validity (including face validity), construct validity (structural, criterion/concurrent, hypothesis testing) and reliability (internal consistency, test-retest). Interpretability is characterized by statistics such as mean and standard deviation which can be translated to a qualitative meaning.

3.3.4 Review methods

3.3.4.1 Search strategy

The search strategy aims to find both published and unpublished studies. A three-step search strategy will be utilized in this review. An initial limited search of MEDLINE and CINAHL will be undertaken followed by analysis of the text words contained in the title and

abstract, and of the index terms used to describe the article. A second search using all identified keywords and index terms will then be undertaken across all included databases. Thirdly, the reference list of all identified reports and articles will be searched for additional studies. Studies published in English will be considered for inclusion in this review. Studies published anytime in the past will be considered for inclusion in this review.

The databases to be searched include:

PubMed

CINAHL

Embase

Cochrane Central Register of Controlled Trials

Emerald Fulltext

MD Consult Australia

PsycARTICLES

Psychology and Behavioural Sciences Collection

PsycINFO

Informit Health Databases

Scopus

UpToDate

Web of Science

The search for unpublished studies will include:

EThOS (Electronic Thesis Online Service), Index to Theses, and Proquest Dissertations and Theses

Initial keywords to be used will be:

collaborat*; collaboration, collaborate, collaborative

multidisciplinary OR transdisciplinary OR interdisciplinary OR multiprofessional OR inter-professional

health*; health, healthcare

measure*; measure, measured, measurement

sensitiv*; sensitive, sensitivity

specificity

instrument

construct

scale

index

valid*; valid, validity, validation

reliab*; reliable, reliability

3.3.4.2 Assessment of methodological quality/critical appraisal

Studies retrieved that meet the inclusion criteria will be assessed for methodological quality by two independent appraisers using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) Checklist (www.cosmin.nl) (see Appendix 1) prior to inclusion in the review. Any disagreements that arise between the appraisers will be resolved through discussion, or with a third reviewer. Currently there is no Joanna Briggs Institute (JBI) appraisal tool that focuses on measurement properties of instruments.

3.3.4.3 Data collection

Data will be extracted from papers included in the review using the COSMIN data extraction tool (see Appendix 1). The reviewers intend to create an Excel spreadsheet of the COSMIN checklist with a four point rating scale, which will be used to record appraisal results and sample characteristics for each measurement property. The data extracted will include specific details about the study quality relating to validity, reliability, interpretability statistics, the sample characteristics (generalizability), study methods and objectives, and outcomes of significance to the review question and objectives.

3.3.4.4 Data synthesis

Effect sizes associated with internal consistency and inter-rater reliability (such as Cronbach's alpha, Cohen's kappa inter-rater scores and/or Kendall's tau) will be reported. If statistical pooling is not possible, the findings will be presented in narrative form including tables and figures to aid in data presentation where appropriate.

This chapter has presented the method of systematic review according to the JBI approach and extended this to the method of systematic review of measurement properties of instruments. The exact protocol used to perform the review has been included here.

In the next chapter the systematic review results are presented. The details regarding the studies including the validation data reported from the studies, the results of the critical appraisal of methodological quality, the levels of evidence of validation and the results of the narrative synthesis of the simple structures of each study are presented.

Chapter 4: Results

This chapter represents the findings of the systematic review conducted to evaluate the measurement of collaboration within healthcare settings. A detailed description of the studies included for critical appraisal is presented followed by the results of the appraisal. This includes search results, validation data extracted from each study, the result of critical appraisal, levels of evidence for validation of each instrument and a narrative synthesis of the simple structures of individual studies.

4.1 Description of studies

Diagram 4.1 represents the results of the database searches and the exclusion of duplicates, irrelevant studies and studies not meeting the inclusion criteria. Initial database searching recovered 2165 unique records which were catalogued in citation management software (EndNote X6). Following removal of duplicate studies, the titles and abstracts were examined and studies were excluded if they were not about healthcare, measurement properties of instruments or not published in English. The final number of studies for additional assessment was 111. After reviewing the full text, studies that were IPE focused, or not about care delivery, or used a sample of less than three participant types, or the sample was inadequately described, or the testing was patient only self reporting were excluded. Furthermore, one study was excluded because the objectives considered the effect of "nesting" individuals within agencies and measuring the perceptions of individuals regarding the collaboration of their agency with other agencies.[157] Even though this study concerned children's mental health agencies, the setting was not care provision but administrative service. This was also the case for the psychometric evaluation of the Collaboration Experience Questionnaire[158] related to the experience of collaboration within collaborative projects. This study was excluded on the basis that it was not care provision. The third study was excluded because the report was an abstract only,[159] leaving 15 studies. Six additional studies were added after examination of references in the articles. A total of 21 studies were included for critical

appraisal representing 12 unique instruments.

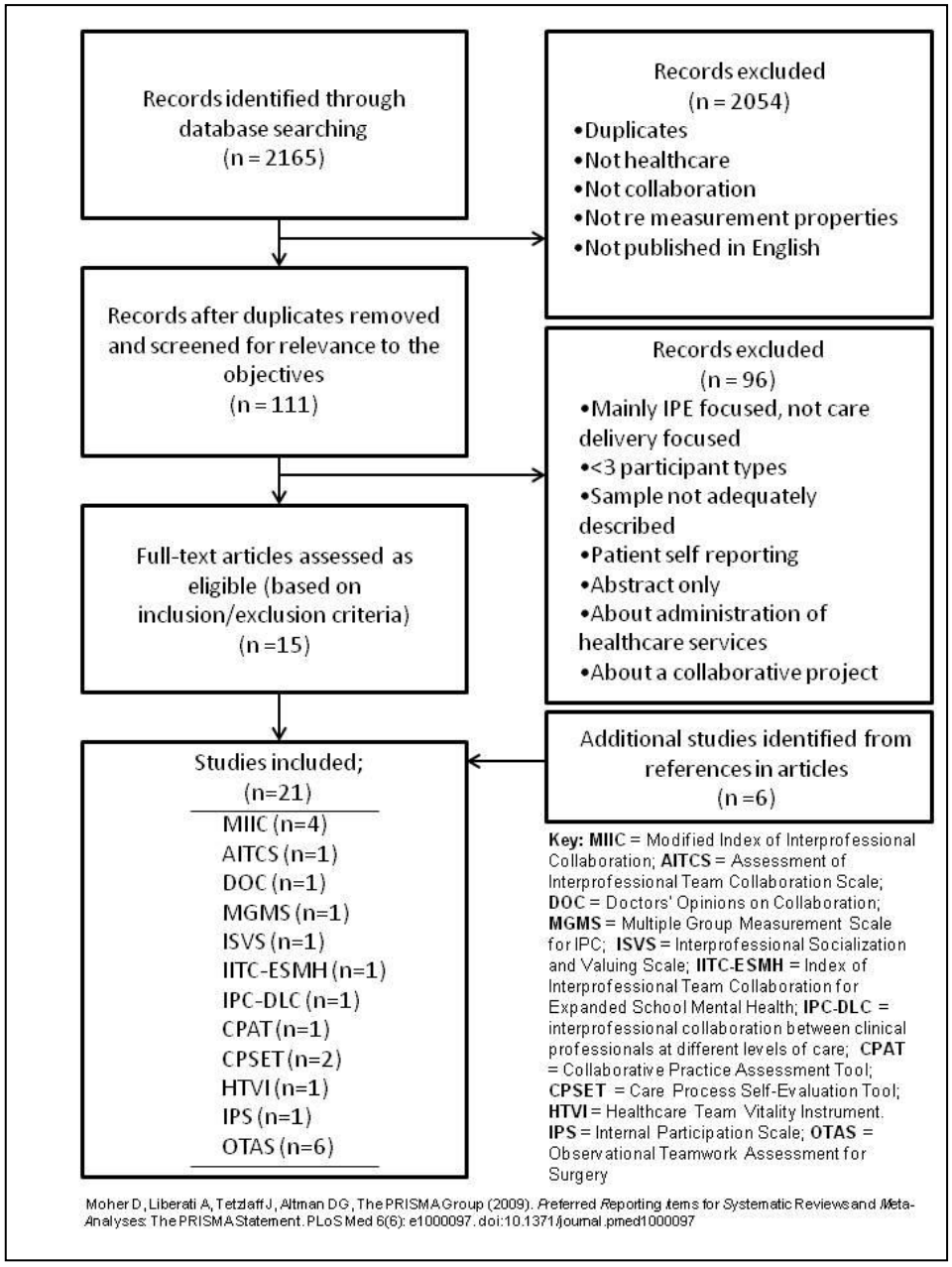

Figure 4.1: Results of literature search and inclusion

Some examples of participant mix included; four professional groups working in school mental health,[152] four groups of medical professions,[160] a mix of medical, paramedical, coordinators and non-professionals (e.g. care logistics),[154] three interdisciplinary groups in surgical teams[85, 99, 109, 111] and both healthcare staff and patients.[161] Studies not reporting the study sample in enough detail, or if the participants represented two participant types or less were excluded, for example nurses and physicians,[162] physicians and surgeons,[163] pharmacists[164] and

general practitioners.[165] Studies of patient self-reporting were excluded, for example Doran et al.,[166] Berendsen et al.[167] and Allen et al.[168]

The studies included were mainly validation studies of measurement instruments that assessed collaboration within the healthcare setting. Most studies reported the development of novel instruments focusing on content validity, structural validity and reliability.[34, 84, 87, 153, 155, 160, 161, 169-171] Some studies evaluated or refined existing instruments.[85, 99, 109, 152, 154] Others contributed to the interpretability of instruments reporting descriptive data such as means, standard deviations and the qualitative meaning of questionnaire scores and statistical treatments.[110, 111, 148-151] Only three instruments, the Modified Index of Interdisciplinary Collaboration (MIIC), the Care Process Self-Evaluation Tool (CPSET) and the Observational Teamwork Assessment for Surgery (OTAS) had more than a single study dedicated to instrument validation. No instruments reviewed in this current study had completed validation studies on all measurement properties.

The settings in which validation data was collected varied and included the Expanded School Mental Health (ESMH) that utilizes interprofessional collaboration to implement learning support and mental health promotion strategies in schools in the U.S.A.;[152] GPs and medical specialists in the Netherlands;[155] hospice teams in the U.S.A.;[148-151] health students in Canada;[153] inpatient wards and services of community and academic hospitals in the U.S.A. and Canada;[170, 171] a mixed sample from acute care, psychiatric and specialized hospitals and primary care organizations from Belgium and the Netherlands;[154] a mixed practitioner sample from a palliative care team; a geriatric assessment team and two family practice teams in Canada;[87] a variety of healthcare teams practicing in Canada;[34] various clinical specialists working in Spanish integrative healthcare organizations;[160] somatic and psychosomatic rehabilitation clinics in Germany;[161] urology surgical teams in teaching hospitals (U.K.); a specialist treatment centre (U.K.); general surgery in a large London teaching hospital (U.K.);[84, 99, 109, 110] surgical procedures performed in Germany[85] and hospital interdisciplinary rounds in a teaching hospital in Chicago

U.S.A.[111]

Some of the studies considered the measurement of personal beliefs, behaviors and attitude in collaboration. For example; how GPs and specialists rate collaboration[155] and the influence on collaborative care,[153] measuring perceptions of collaboration,[150, 151] and nurse, allied health and physicians judgments of IPC.[171] Other studies validated instruments measuring care process organization,[154, 169] assessing collaborative relationships,[34] measuring IPC between clinical professionals at different levels of care,[160] assessing team function [84, 85, 87, 99, 109-111, 152, 170] and measuring internal participation (a core component of patient-centeredness teamwork) in both healthcare staff and patients.[161]

A distinction was made between healthcare provision, for example hospice, hospital or community HCS[149, 152, 153] as opposed to studies of interprofessional educational (IPE), for example online IPE in a dementia case study for health science students.[172] Because generalizability of tools to HCS was central to this review, studies of IPE were excluded. A table of all excluded studies (n=96) and the reasons for exclusion can be found in Appendix 4.

Characteristics of included studies (see Appendix 5) shows the characteristics of each study; a summary of the findings relating to the methodological quality of the assessed measurement properties for each study and a summary of the instrument's utility and interpretability. The *sample characteristics* indicate the populations to which each instrument may be generalizable and the *setting* relates to the healthcare environment from which the sample has been recruited.

4.2 Review finding/results

4.2.1 Methodological quality

The papers appraised in this review are diverse in their theoretical underpinnings, target population and measurement objectives. Methodological quality is highlighted in this section and substantive findings are presented for each instrument. The result of the appraisal of methodological quality is summarized in Table 4.1. The coding system used to present the data

is represented in Levels of Evidence table (see Table 4.2).

It is essential for the reader of this review to understand an important distinction when interpreting the results presented here. The *validity or reliability statistic* such as alpha, kappa or an intraclass correlation coefficient reported in an individual included study is not the same as the *level of evidence* for validity or reliability. The level of evidence relates to the methodological quality of one or more studies contributing to the overall validation of the instrument. Say for example, a study of internal consistency produces a result of alpha = 0.90, which indicates good interrelatedness. But if the sample size is n=5 and only one study has been performed, the level of evidence for reliability is 'unknown'. If however three studies have been performed all with n>30, and consistent results were produced between studies, the level of evidence may be 'strong'. Therefore, when considering the results presented here the reader is advised to reference Table 4.2 to interpret the specific meaning of the terms assigned to the levels of evidence and the criteria upon which these terms are ranked.

It was possible to evaluate the measurement objectives stated in each of the individual study reports and on this basis create generalized categories under which we could report our findings. Objectives in measuring collaboration identified in the individual studies included: beliefs, behaviors and attitudes (measuring the socio-cultural aspects of the healthcare setting); between different levels of care (for example measuring collaboration between primary care practitioners and hospital specialists); in multi-rater on target group (how professional groups rate collaboration with other professional groups for example how nurses rate collaboration with doctors); of perceptions (based on how an individual mentally constructs their experience of collaboration); relationships (measuring interpersonal factors associated with collaboration such as cooperation and partnership); in assessing teams (using collaboration as a measure of team function) and measuring internal participation (inter-professional patient centred teamwork) of both healthcare staff and patients.

Table 4.1 Results of the critical appraisal of methodological quality per questionnaire (refer to Table 4.2 for interpretation).

	Reliability		Validity				
Instrument	Internal consistency	Reliability	Content validity	Structural validity	Hypothesis testing/concurrent validity	Criterion validity	Cross cultural
MIIC	+	na	++	na	na	?	na
AITCS	?	na	+++	?	na	na	na
DOC	+	na	+	+	+	na	
MGMS	++	na	+++	++	++	++	na
ISVS	+	na	+++	+	na	na	na
IITC-ESMH	+++	na	+++	+++	na	na	na
IPC-DLC	+	na	+++	+	na	na	na
CPAT	?	na	+++	?	na	na	na
CPSET	+++	++	+++	+++	na	+++	na
HTVI	na	na	+	+	na	+	na
IPS	++	na	?	++	na	++	na
OTAS	na	++	+++	na	+	na	+

na = not assessed

Key: MIIC = Modified Index of Interprofessional Collaboration; AITCS = Assessment of Interprofessional Team Collaboration Scale; DOC = Doctors' Opinions on Collaboration; MGMS = Multiple Group Measurement Scale for IPC; ISVS = Interprofessional Socialization and Valuing Scale; IITC-ESMH = Index of Interprofessional Team Collaboration for Expanded School Mental Health; IPC-DLC = interprofessional collaboration between clinical professionals at different levels of care; CPAT = Collaborative Practice Assessment Tool; CPSET = Care Process Self-Evaluation Tool; HTVI = Healthcare Team Vitality Instrument; IPS = Internal Participation Scale; OTAS = Observational Teamwork Assessment for Surgery

Table 4.2 Levels of Evidence (according to Paalman et al.)[124(p4)]

Level	Rating	Criteria
Strong	+++ or ---	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
Moderate	++ or --	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
Limited	+ or -	One study of fair methodological quality
Conflicting	+/-	Conflicting findings
Unknown	?	Only studies of poor methodological quality

4.2.2 Measuring collaboration beliefs, behaviours and attitudes (two studies/two instruments)

The Interprofessional Socialization and Valuing Scale (ISVS) was designed "...to capture the beliefs, behaviors, and attitudes of professionals that influence and are influenced by their transactions in enacting collaborative care approaches".[153(p79)] The study used a sample of clinical kinesiologists, dieticians, medical doctors, nurses, occupational therapists, physical therapists, psychologists, social workers, speech pathologists and others (composition not described). Excellent methodological quality produced strong evidence for content validity. The authors are experts in their field and developed a comprehensive set of items that were evaluated by a professional working group. The items are worded for professionals and the study sample reflected this. This study demonstrated good internal consistency (alpha for whole scale 0.90 and subscales 0.79-0.89), however the evidence for internal consistency was limited because it was unclear how missing items were handled. Factor analysis produced a three factor solution that accounted for 49% of the variance. The three factors were labelled self-perceived ability to work with others, value in working with others and comfort with working with others. Methodological quality suffered from a lack of reporting of missing items, therefore there is limited evidence for structural validity. In summary the ISVS has application in HCS for the

assessment of socio-cultural aspects of collaborative practice of professionals and has limited to strong evidence of validity. The ISVS has the potential application in settings where the team is working towards developing comfort in working together. Even though one item asks about the importance of having patient and family as "members of a team", the questionnaire does not accommodate the capacity to include these members as a part of the data collection. The generalizability of this study is limited due to the sample consisting predominantly of occupational therapy and nursing students. Further research is needed using a broader sample to readdress the validity and reliability properties of this tool.

The Doctors Opinions on Collaboration (DOC) questionnaire study[155] used a sample of physicians, surgeons and support specialists and included psychiatrists, internists, paediatricians, cardiologists, neurologists, rehabilitation doctors, pulmonologists, dermatologists, clinical geriatricians, allergologists, rheumatologists, ophthalmic surgeons, gynecologists, general surgeons, urologists, orthopaedic surgeons, orofacial surgeons, plastic surgeons, ENT doctors, thoracic surgeons, support specialists, radiologists, radiotherapists, microbiologists and pathologists. The content of the DOC questionnaire is to measure the beliefs and attitudes of GPs and specialists towards the collaborative relationship with each other and was developed from previous qualitative research,[173, 174] input from key experts and from the evaluation of a pilot study. There was limited evidence of content validity as the methodological quality suffered because it was not clear as to the theoretical basis of the instrument. Factor analysis produced a five factor solution accounting for 55% of the variance. The factors were labelled organization, communication, professional expertise, image, and knowing each other. There was limited evidence for structural validity. Adequate internal consistency (alpha for subscales 0.64-0.83) was demonstrated however, there was limited evidence for reliability due to a lack of detail regarding missing items. Four hypotheses developed from the author's qualitative studies were tested. The results of the comparisons between groups were consistent across each of the qualitative study findings. However, evidence for validity was limited due to a lack of detail

regarding missing items. In summary, for assessing collaboration between doctors and specialists the DOC questionnaire is a useful tool with limited evidence of validity. It is unlikely to be applicable to measuring collaboration in situations where the participant mix is more complex beyond GPs and specialists.

4.2.3 Measuring collaboration between different levels of care (one study)

Nuno-Solinis et al.[160] developed the IPC-DLC (InterProfessional Collaboration between two Different Levels of Care) to assess clinicians perception of collaboration between two levels of care (primary and specialized) and was developed with a sample consisting of primary care nurses, GPs, paediatricians, hospital specialists and hospital nurses from various primary care and specialized care disciplines. The authors based content development on a strong theoretical model (Danielle D'Amour)[28] and responded with item amendments after the evaluation by five experts in care integration as well as Danielle D'Amour, providing strong evidence for content validity. Exploratory factor analysis produced a two factor solution explaining 59.3% of the variance with good fit. The two factors comprising collaboration were labelled as interpersonal relationships and organizational characteristics. There was limited evidence for structural validity as it was not clear how missing items were handled. This study demonstrated good internal consistency (alpha >0.80) with limited evidence for reliability as it was not clear how missing items were handled. The authors note that it would be beneficial to test this instrument "... using samples from other groups of healthcare professionals and organisational contexts". Justifiably this tool has limited generalizability and limited validity. However it may be useful in measuring collaboration in a sample of doctors and nurses from different levels of healthcare.

4.2.4 Measuring collaboration in multi-rater on target groups (one study)

A major focus of the Multiple Group Measurement Scale (MGMS) for IPC[171] was to develop an instrument to measure collaboration between multiple groups consisting of three

participant types; nurses, physicians and allied health practitioners. A round-robin method was used to collect data and allowed a multi-rater on target group assessment. In the report background the authors discuss measurement equivalence invariance (ME/I) when assessing collaboration in multiple clinical participant types. ME/I addresses the question of whether respondents to questionnaires interpret a measure in a conceptually similar way or do raters define performance the same way when rating the same target on identical performance dimensions.[175] However, the developers did not take the additional step of investigating ME/I. They posit that an instrument should be valid in at least one rater group. In this study nurses are considered central to the healthcare setting and the content of the questionnaire is adapted from a well known questionnaire, the Collegial Nurse-Physician Relations Subscale of the Nursing Work Index (NWI-NPRS).[176]

The NWI-NPRS is an adaptation of a gold-standard questionnaire in nursing collaboration assessments and evidence for content validity of the MGMS was strong. Evidence for internal consistency was moderate and this study demonstrated good reliability reported as Raykov's composite reliability statistic[177] across all three professional groups that ranged from rho=0.71 to rho=0.88. The NWI had a reliability statistic of rho=0.92. There was moderate evidence for inter-rater reliability estimated between participant's responses (hospital level). Average inter-item correlation across hospital sites was variable with an average of 0.59 which is below the acceptable value of 0.60.

Moderate evidence supported structural validity for a three factor solution by exploratory and confirmatory factor analysis. The three factors were labelled communication, accommodation and isolation. The MGMS is similar to the NWI-NPRS and the subscales of the Attitudes Toward Health Care Teams Scale(ATHCTS).[178] Moderate evidence supported criterion validity. Correlation with the NWI-NPRS showed high interclass correlation coefficients with the three factors of the MGMS as was hypothesized. Also, as hypothesized, the MGMS showed low correlation with the ATHCTS. These results showed a clear conceptual

difference between the MGMS and the ATHCTS and a conceptual similarity with the NWI-NPRS; confirming this was a low correlation between the NWI-NPRS and the ATHCTS. Furthermore, it was hypothesized that the differences of the mean scores for the MGMS and the mean scores for the NWI-NPRS (criterion) should correlate between hospital sites. This was confirmed with the nurse's assessment of physicians.

The MGMS demonstrates potential as a valid and reliable tool (moderate to strong evidence) for the rating of collaboration of physicians by nurses. Further validation is needed to extend this tool's application to a broader participant base.

4.2.5 Measuring perception of collaboration (two studies/one instrument)

Based on Bronstein's model of collaboration and the Index of Interdisciplinary collaboration (IIC),[27] Oliver, Wittenberg-Lyles and Day modified the IIC (MIIC) and tested its validity, reliability and interpretability in a sample of nurses, social workers, chaplains, administrators, and other clinicians in the hospice setting.[150] Content validity was assumed by the authors to be similar to the original IIC and was additionally assessed by five hospice workers, providing moderate evidence of face validity. The MIIC showed internal consistency (alpha of 0.935 for the total instrument and between 0.767 and 0.867 for the subscales) and there was limited evidence to support this due to no factor analysis and the study only referring to the earlier development.[27] Means and standard deviations were reported. The authors stated the MIIC is comparable psychometrically to the original instrument, and a paired sample t-test was performed that showed no significant statistical difference between the two instruments, though these results are not in the report. Therefore, evidence for criterion validity is limited.

Oliver, Wittenberg-Lyles and Day[151] used the IIC in its modified form (MIIC) to determine variance in perception between different hospice programs and between nurses, social workers, chaplains, administrators, and other clinicians in those settings. This study contributed interpretability of the MIIC and reported means and standard deviations.

Cronbach's alpha was calculated for the total scale and each subscale, however the study does not report this result and the authors refer to their prior validation study (see above).[150] Variance in perception was determined using ANOVA. The results indicated there was no significant difference in perception of collaboration between participant types, but there was a significant difference between hospice programs.

In summary, these two studies contribute to the interpretability of the MIIC and its generalizability to the hospice setting comprising of a complex mix of participant types. The capacity of the MIIC to determine differences in the measured perception of collaboration between hospice programs suggests this tool may be useful to gauge the effect of implemented programs aimed at improving collaborative care. Studies to determine responsiveness validity of the MIIC are needed.

4.2.6 Measuring collaborative relationships (one study)

The Assessment of Interprofessional Team Collaboration Scale (AITCS) was developed from the need to assess collaborative relationships in HCS, especially as most existing tools focus on team effectiveness and lack a patient focus regarding roles within the collaborative team.[34] The participants in this study were registered nurses, medical practitioners, physiotherapists, occupational therapists, pharmacists, social workers, dieticians and practice nurses.

The developers based the AITCS on Sullivan's model of collaboration involving the process of creating power-sharing relationships.[26] Four key collaborative domains evolve from this model; coordination, cooperation, shared decision making, and partnership. Based on this model, 24 IPE experts were contacted to review the items for face and content validity. The result was a 47 item questionnaire supported by strong evidence for content validity.

Factor analysis and refinement using confirmatory factor analysis revealed a three factor solution accounting for 61.02% of the variance. The three factors were labelled coordination, cooperation and partnership/shared decision making. Evidence for structural validity remains

unknown due to a small sample size (n=122-123, items=37) which is less than five times the number of items (<5*#items). Test for internal consistency demonstrated an alpha for the whole instrument=0.98 and subscales 0.80-0.97. Evidence of reliability is unknown due to the small sample size. High alpha values suggested the possibility of some redundant items.

In summary the AITCS is an instrument with strong evidence of content validity only that addresses collaborative relationships in the healthcare setting and has a potential for use with collaborative teams with a diverse participant type mix. The AITCS needs further study to evaluate reliability and it may also be possible to reduce the total number of items.

4.2.7 Measuring collaboration in assessing teams (thirteen studies/six instruments)

The Collaborative Practice Assessment Tool (CPAT)[87] was designed for the purpose of diagnosing the relative strengths and weaknesses of individual teams within HCS. The developers of the CPAT rejected other tools identified in the literature because those tools did not include some critical concepts to collaborative practice such as the role of the patient/family. In the first phase pilot study, CPAT was developed from a review of concepts identified from the literature and with expert opinion; there was strong evidence for content validity. The questionnaire was refined with EFA to produce 42 items with a seven factor solution. A second pilot used 111 participants including practice nurses, medical practitioners, physiotherapists, occupational therapists, social workers, pharmacists, dieticians, housekeepers, porters, spiritual care, clerical staff and others and CFA to produce an 8 factor solution with satisfactory goodness-of-fit indices. The factors were labelled mission, meaningful purpose, goals; general relationships; team leadership; general role responsibilities and autonomy; communication and information exchange; community linkages and coordination of care; decision-making and conflict management; and patient involvement. The evidence for structural validity remains unknown due to the small sample size (<5*#items). Internal consistency was good (alpha for subscales 0.73-0.84), however evidence of reliability is unknown due to the

small sample size.

The CPAT has utility in the healthcare setting with a complex mix of participant types to assess collaboration as an assessment for team function. Evidence of validity is unknown to strong and further validation of this tool is needed especially responsiveness if the tool is destined to be used to assess team weakness and strength and logically to retest after any intervention and time period.

Two studies were included that validated the Care Process Self-Evaluation Tool (CPSET). The original study by Vanhaecht et al.[169] identified the need for a tool that assesses how clinical pathways influence the process of care. Of theoretical importance to CPSET development is the Realistic Evaluation-CMO Framework(CMO).[179] CMO means context+mechanism=outcome. 'Realist evaluations ask not, "What works?" or "Does this program work?" but ask instead, "What works for whom in what circumstances and in what respects, and how?" [179(p3)] CPSET content development was extensive including multiple focus groups, a convergence phase to group candidate topics under the CMO stated as items, cognitive testing and finally a Delphi study to rank the importance of items and reduce the number of items. A further face validity study using a participant mix of clinical pathway facilitators, medical doctors, allied health professionals, senior hospital managers, supporting departments, primary care, nurses and patients produced a 'beta version' of the CPSET with 24 context items, 51 mechanism items, nine outcome items and three general CMO items. This developmental process produced strong evidence for content validity. Structural validity using a split sample for EFA and CFA included 511 returned surveys from medical doctors, nurses, allied health professionals and pathway coordinators and resulted in 29 items and a five factor solution accounting for 65% of the variance. The five factors were labelled patient-focused organizations, coordination of care, communication, collaboration with primary care and follow-up of care. There was strong evidence for structural validity and for criterion validity. An analysis of the CPSET five factors with the three CMO questions produced statistically

significant Kendall *T* item correlations. Internal consistency was demonstrated for the five factors (alpha 0.776-0.928) supported by strong evidence of reliability. There was moderate evidence for inter-rater reliability estimated using intraclass correlations, which were all statistically significant (ICC 0.280-0.704).

Seys et al.[154] further evaluated the CPSET and complied a cut-off score table using a sample of nurses, medical doctors, paramedics, coordinators and others (e.g. care logistics). Creating this table from CPSET scores is proposed to be helpful for healthcare managers to rank teams in their facility. Strong evidence of structural validity from confirmatory factor analysis on 3139 questionnaires showed good fit for the collected data. Internal consistency was good (alpha 0.869-0.950) and there was strong evidence for reliability. This study went some way to adding to interpretability of the CPSET. A multilevel analysis at 'team' and 'hospital' levels showed the interclass correlation coefficients of the scores of teams were higher than hospitals, indicating less variance within teams than within hospitals. Comparisons of CPSET scores between demographic factors in team members showed statistically significant differences between groups. In summary, the CPSET proves to be a reliable and valid instrument (strong evidence) for evaluating care processes, with a significant component being team collaboration.

The Healthcare Team Vitality Instrument (HTVI)[170] was developed to assess "team vitality" in healthcare teams working in hospital units using a two phase process with a sample of nurses, physiotherapists, assisting personnel, unit secretaries, physicians and others (composition not described). An initial literature review identified the common desirable domains and was the basis for the HTVI. Cognitive interviews of 18 participants (15 were nurses) tested content validity. However, evidence of content validity is unknown because it was not clear if all the items together reflected the constructs being measured. There was limited evidence due to lack of reporting regarding missing items to support structural validity, which included EFA and CFA that produced a 4 factor solution explaining 58% of the variance with

adequate goodness-of-fit. The factors were labelled support structures; engagement and empowerment; patient care transitions; and communication. Responses to the HTVI items were strongly to moderately correlated with responses to the NWI, NWI-R and the Agency for Healthcare Research and Quality (AHRQ), but with limited evidence for criterion validity due to lack of reporting regarding missing items. Further research should seek to add an assessment of this instrument's reliability. The generalizability of this instrument is limited due to the participants of this study being mostly nurses. In summary, the HTVI is a tool with limited evidence of validity and further studies will be valuable. Regardless, the HTVI may be useful in measuring team collaboration in hospital units.

The Index of Interprofessional Team Collaboration for Expanded School Mental Health (IITC-ESMH)[152] was developed using the Bronstein model[27] of collaboration explicitly for measuring the functioning of interprofessional teams in Expanded School Mental Health (ESMH) settings using a sample of nurses, psychologists, social workers, counsellors and others (composition not described). According to Mellin et al. "ESMH utilizes interprofessional collaboration to implement learning support and mental health promotion strategies in schools".[152 (p515)] Evidence of content validity was strong; the IITC-ESMH was developed by 21 geographically diverse experts in ESMH who were consulted to establish content validity and pilot test the instrument. Internal consistency was good (alpha 0.49-0.91) which was supported by strong evidence of reliability. Evidence for structural validity was strong and the EFA using principal component analysis with promax rotation produced a four factor solution (consistent with Bronstein) that accounted for 63.25% of the variance. Further CFA with a new data set is needed as well as reliability and criterion validity assessments. In summary, this study is a significant contribution to the validity of the IIC/Bronstein model and is useful in measuring team function in ESMH.

Two studies by Wittenberg-Lyles et al.[148, 149] used the MIIC (which contributes to the interpretability of the tool in the hospice setting) with samples of nurses, social workers,

chaplains, administrators and others (composition not described). The hospice setting has been described "...as an exemplar for other geriatric healthcare teams..." in holistic and an interprofessional approach to healthcare.[148] The first study[148] used a mixed method approach by observing team meetings (qualitative) and administering the MIIC (quantitative). This type of research approach proved insightful as the qualitative findings indicated that collaboration occurred outside of the team and between other teams outside of the hospice.

The second study[149] used the same approach and reported on differences in collaborative acts verses perceptions of collaboration in the hospice setting with nurses, social workers, chaplains, volunteer coordinators, bereavement coordinators, medical students, home health aides and an executive director of the hospice. "[The] team's reflection on process was... the most demonstrated collaborative act... yet it was perceived... as the least collaborative act".[149(p7)] When caregivers were present in team meetings reflection on process dropped from the highest to lowest collaborative act in meetings. Wittenberg-Lyles et al. concluded that "...perceptions of interdependence and flexibility were much higher than enacted collaborative practices in IDT meetings regardless of caregiver involvement".[149(p7)] As demonstrated in these two studies, the MIIC has the potential to reveal valuable insights into team function through the measurement of collaboration.

The Observational Teamwork Assessment for Surgery (OTAS) tool differs from the other instruments evaluated in this current study. Unlike the other instruments which are questionnaires with the data collected relating to each individual participant's perception, the OTAS is a checklist that records an observer's identification of collaborative acts in the operating room (OR). OTAS consists of a procedural checklist and a behavioural checklist. The procedural checklist records the surgical team's performance in the critical tasks domain therefore this component of the OTAS is not relevant for this systematic review. The behavioural checklist identifies the behavioural dimensions of teamwork in preoperative, intraoperative and postoperative phases of surgery across three sub-teams including surgical,

nursing and anaesthesia.

The OTAS is based on theoretical constructs of collaboration consistent with those identified in this current review. In 2007 Undre et al.[180] developed the OTAS based on Dickenson and McIntyre's Model of Teamwork.[181] This model identifies the dimensions of communication (quality and the quantity of the information exchanged among members of the team), coordination (management timing of activities and tasks), cooperation/backup behavior (assistance and support of other team members and correcting errors), leadership (provision of directions, assertiveness and support among members of the team) and monitoring/awareness (team observation and awareness of ongoing processes).[180]

Hull et al.[84] conducted a content validation of the OTAS instrument by using a two phase process to test two hypotheses. Hypothesis 1; blinded observers will demonstrate interrater agreement for exemplar behaviors. Hypothesis 2; exemplar behaviors will show content validity via expert consensus. Exemplar behaviors (exemplars) are defined as "... key observable behaviors that indicate exemplary teamwork and are associated with effective, safe surgical practice".[84(p235)] In Phase 1, data was collected from the observation of 30 general surgical procedures by two blinded observers (psychologists) in a London teaching hospital. Correlational analysis (intraclass correlation coefficients) of the results determined the strength and direction of the relationship between the two rater's scores and frequency analysis (Cohen's kappa and percentage agreement) determined exemplar observability and interrater agreement. The results showed strong correlations for all exemplars (ICC 0.64 – 0.77, $p<0.001$) and there was high interrater agreement (kappa $\geq$ 0.41, percentage agreement $\geq$ 70%) for 109 of 130 exemplars. This enabled the determination of problematic exemplars. In Phase 2, 56 problematic and new exemplars were refined via expert consensus. The expert panel consisted on 15 experts (5 surgeons, 5 nurses, 5 anesthesiologists). The experts rated the exemplars, producing a content validity metric (CVM) to evaluate the exemplar contribution to teamwork and safety and to construct a rank order for exemplars. Three experts then clarified or removed

exemplars to complete the process. The methodological quality for assessing the reliability of the OTAS instrument in this study was fair and for content validity was excellent.

Interrater reliability of the OTAS was tested by Undre et al.[109] for application in urological surgery. This study was conducted over 50 urological procedures observed in two operating theatres in one teaching hospital and one specialist treatment centre in the U.K., which showed adequate agreement between two observers via correlating the two observers behavioural ratings (r=0.35 to r=0.72, $p<0.05$). The methodological quality for assessing reliability was fair.

Passauer-Baierl et al.[85] adapted the OTAS instrument for ORs in Germany. The aims of the research were to "... translate, adapt, and refine a German version of the OTAS (OTAS-D) as well as test its face validity, applicability, and interrater reliability".[85(p306)] The researchers used a systematic translation and adaption process, expert content validation aiming to establish functional equivalence and to test reliability to establish metric equivalence. The standardized translation process involved a forward-backward translation and revisions for clarity. The next phase included an expert panel of nine OR professionals (3 surgeons, 3 nurses, 3 anesthesiologists) undergoing semi-structured interviews relating to the OTAS-D exemplar behaviors and the relation to team functioning and patient safety in German ORs. The interviews were analysed using qualitative thematic content analysis. This resulted in OTAS with 115 exemplars relevant to German ORs. The methodological quality for assessing cross cultural validity was fair and good for content validity.

A test for reliability was conducted using 2 independent raters (blinded) over 11 randomly selected surgical procedures. Interrater agreement was almost perfect (kappa $\geq$ 0.80) for 57 of 115 exemplars. Other exemplars showed adequate agreement (kappa 0.20-0.79) with 25 exemplars showing problematic agreement (kappa < 0.20) needing revision. Furthermore, this study evaluated the reliability of the OTAS scoring between two blinded observers using intraclass correlations (ICC). All OTAS behavioural ratings shows statistically significant

agreement between raters and the global ICC was excellent (0.80, $F(485)= 9.21$ $p< 0.001$). Further evaluation of inter-rater agreement by calculating the frequency of inconsistent rating between observers by more than one scale category showed a 98.36% consistency. The methodological quality for assessing reliability was fair.

The study by O'Leary et al.[111] utilized the OTAS to assess teamwork during structured interdisciplinary rounds on medical units in a tertiary care teaching hospital in Chicago U.S.A. This represents an adaptation of the OTAS for a novel application. Although this study was not primarily a validation study, it adds interpretability data for the OTAS tool in addition to reliability data. Appropriately, the study tests the OTAS for interrater reliability. Spearman's rank correlation coefficient was used to show that interrater reliability across different medical units was excellent (rho=0.75) and across sub-teams was good to excellent (rho=0.53-0.76), with the physician sub-team showing poor interrater reliability (rho=0.35). However, the methodological quality for assessing reliability generated by this study was poor due a small sample size (n<30). This study demonstrated interpretability of the OTAS instrument by identifying differences in performance across units, domains and subteams.

Sevdalis et al.[99] addressed the construct validity of the OTAS instrument by using 12 urological procedures conducted in two London teaching hospitals by applying a hypothesis test. The methodological approach relied upon an established method that is based on the premise that an instrument should be sensitive enough to capture differences between novices and experts.[182, 183] If two experts agree as much as an expert and a novice then instrument is not measuring the underlying behaviors or the tool is initially unnecessary.[99] Data was paired between Expert 1 and Expert 2 and between Expert 1 and Novice. Analysis using Pearson's *r* correlation coefficients to determine the strength and direction of ratings between experts and novice was performed. Analysis of Variance (ANOVA) was used to determine statistically significant differences between expert and novice ratings. The correlations provided evidence to support the hypothesis that ratings for Expert 1 and Expert 2 correlated more strongly than

rating between Expert 1 and Novice (H_{1a}). Furthermore, by determining the differences in mean scores it was demonstrated that Expert 1 and Novice ratings showed sizable and significant inconsistencies in 11 of 15 rated behaviors. In contrast, only three of 15 rated behaviors were significantly different between Expert 1 and Expert 2, supporting the hypothesis that more numerous and sizable discrepancies were seen between Expert 1 and Novice than between Expert 1 and Expert 2 (H_{1b}). The authors concluded that the "...findings... suggest that OTAS demonstrates construct validity". [99(p1049)] This study produced fair methodological quality for assessing construct validity because raters were from a narrow sample of two experts and one novice. The authors suggested replicating the results with a larger range of disciplines.

The utility of the OTAS instrument has been demonstrated by Russ et al.[110] in a study evaluating the effects of a short term training program for clinical and non-clinical novice assessors rating teamwork in the operating room. The ability of novice assessors to reliably use the OTAS after short term structured training and the ability of novice assessors from different professional backgrounds (2 psychologists and 2 surgeons) to use the OTAS reliably was the aim of this research. Data was collected during 14 general surgical procedures in a large teaching hospital in London U.K. Descriptive statistics (mean/SD) for experts and novices, intraclass correlation coefficients, Pearson's *r* (transformed to *Z* scores for each OTAS behavior and subjected to ANOVA) were calculated. *Z* scores were used to test differences in the learning curves between surgeon and psychologist novices. This study demonstrated excellent interpretability using the OTAS, showing acceptable interrater reliability between experts and novices at the end of training (ICCs $\geq$ 0.68), improved calibration across the 10 observed cases, an observed ceiling effect for the calibration of coordination, and no significant difference between surgeons and psychologists in calibration with the expert.

In summary, the OTAS is a valid and reliable instrument for the assessment of collaboration within surgical teams. There is moderate evidence of reliability, strong evidence of content validity, limited evidence of construct validity and limited evidence of cross-cultural

validity. Future research using the OTAS in varying surgical and non-surgical scenarios should add further evidence for the validity of this useful and potentially adaptable instrument.

4.2.8 Measuring internal participation (one study/one instrument)

Internal Participation (IP) is defined as "... teamwork between two or more healthcare professionals from different disciplines to provide comprehensive services to patients, or in other words interprofessional patient-centred teamwork".[161(p375)] The theoretical model of patient centred interprofessional participation proposed by Korner[161, 184] describes both internal and external participation where internal participation occurs between the healthcare professional and the team and external participation occurs between the patient and the team or the patient and a healthcare professional. Therefore the IPS is an instrument designed to measure internal participation from both the staff's perception and the patient's perception.

The IPS was developed from a theoretical base[184-186] as well as incorporating task specific and social elements of team functioning models.[187, 188] However, evidence for content validity of the IPS is unknown as there has been no assessment of the items' relevance to the target population with only an assessment of appropriate language by a mixed group of three healthcare research experts, two rehabilitation experts and three patients.

The internal consistency of the IPS was good (alpha ranging between 0.871 and 0.878) with moderate evidence of reliability. Inter-item correlations ranged from 0.377 to 0.733 for patients and 0.349 to 0.686 for staff. Evidence for structural validity was moderate. Exploratory factor analysis showed all items loaded on one factor which explained 61.1% of the variance for staff and 62.3% for patients. Confirmatory factor analysis showed a good model fit and explained 54% of the variance. The six items were allocated short labels consistent with the central aspects of collaboration theory and these were climate, cooperation, agreements, coordination, communication and respect.

The IPS was tested for discriminate validity against the SDM-Q-9[189] or the SDM-Q-Doc[190] and the IRES-24[191] questionnaires. Convergent (criterion) validity of the IPS was tested

against the Questionnaire on Staff Satisfaction in Medical Rehabilitation[192] and the Questionnaire on Patient Satisfaction.[193] There was moderate evidence of criterion validity for the IPS, which showed high correlation (r=0.593, p≤0.001) for patient satisfaction and high correlation (r=0.551 to 0.748, p≤0.001) for staff satisfaction.

In summary, the IPS presents as a short instrument validated for the use in measuring internal participation in medical rehabilitation settings from both the healthcare professional and the a patient's perception. Further validation needs to provide evidence of content validity, reliability (inter/intrarater) and responsiveness.

4.3 Synthesis of latent variables

The purpose of factor analysis is to reduce complex variability data to a smaller number of factors. The outcome of factor analysis is sometimes referred to as *simple structure*. In this review it became possible to extract the factors identified in each study that conducted factor analysis (10 studies) and perform a narrative synthesis of this information. The process of extracting the factors involved; examining the factor descriptors in each study, as well as referring to the items in the instruments for clarity and allocating the factor descriptor to a table with general headings aligned with theoretical statements from Bronstein,[27] D'Amour,[28] Sullivan[26] and San Martín-Rodríguez[29] (see Table 4.3). The MIIC inherits the factor structure determined by Bronstein.[27] The IPS study produced a single factor structure (internal participation), and the six items of this instrument were labelled with summary descriptors which we used in the synthesis. The OTAS instrument, being a checklist and not a questionnaire, did not undergo structural validation via factor analysis. Therefore, the latent variables of the OTAS are defined by the behavioral constructs of communication, coordination, cooperation/backup behavior, leadership and monitoring/situational awareness.[84] These constructs were used in the synthesis (see Table 4.3).

Nine general summary factors were extracted from the synthesis. We have given these general summary factors descriptors to best encapsulate the results and these are;

- Organizational settings, support structures, purpose and goals
- Communication
- Reflection on process
- Cooperation
- Coordination
- Role interdependence and partnership
- Relationships
- Newly created professional activities
- Professional flexibility

Table 4.3 Narrative synthesis of factor structures of each instrument

Instrument	Role inter-dependence & partnership	Organisational support structures, purpose & goals	Communication	Cooperation	Coordination	Reflection on process	Relationships	Newly created professional activities	Professional flexibility
Theory	Creating effective working relationships[25] Value the contribution of others and common goals[21] interactions where all are dependent on the others to reach the goals[23]	The organization's structure, philosophy, administration, resources and coordination mechanisms[24] Shared responsibility in the process of reaching goals[23]	Willingness to communicate[24]	Contribution of views and valuing those of other team members[25]	Achieving mutual goals by working together[25]	Attention to the process of working together[23]	Collegial relationship that involves open communication, mutual respect and trust[21] Trust… and mutual respect[24]	New activity and services not achieved without collaboration[23] The whole is greater than the sum of its parts[21]	The deliberate occurrence of role blurring[23]
MIIC	Role Interdependence	Collective ownership of Goals				Reflection on Process		Newly Created Professional Activities	Professional Flexibility
AITCS	Partnership/shared decision making			Cooperation	Coordination				
DOC	Knowing each other		Communication		Organisation†	Professional expertise	Image‡		
MGMS	Isolation		Communication	Accommodation					
ISVS	Value in working with others			Self perceived ability to work with others			Comfort in working with others		
IITC-ESMH	Role Interdependence	Collective ownership of Goals				Reflection on Process		Newly Created Professional Activities	Professional Flexibility
IPC-DLC		Organisational settings					Personal relationships		
CPAT	General role responsibilities, autonomy,	Mission, meaningful purpose, goals	Communication and information exchange	Decision-making and conflict management	Community linkages and coordination of care	team leadership§	General relationships		Patient involvement¶

CPSET		Patient-focused organization	Communication with patient and family	Collaboration with primary care	Coordination of the care process	Monitoring and follow-up of care process	
HTVI	Engagement/ empowerment	Support structures	Team communication		Patient care transitions		
IPS	Agreements	Climate	Communication	Cooperation	Coordination		Respect
OTAS		Leadership*	Communication	Cooperation	Coordination	Monitoring and situational awareness	

The next chapter discusses the implications of the results obtained. The importance of measuring collaboration is placed in the context of an understanding of complexity and the biopsychosocial approach to healthcare. Issues surrounding measurement relating to simple structure, the value of triangulation, measurement equivalence and the potential of narrative synthesis of simple structure are discussed.

Chapter 5: Discussion

In this chapter a discussion of the results precedes a consideration of the importance of addressing complexity in measuring social phenomena such as collaboration within the healthcare setting. The use of factor analysis and its limitations in addressing complexity is discussed as is the importance of using a biopsychosocial paradigm to capture some of the complexity relating to healthcare. The use of a narrative synthesis of simple structure is discussed. Finally the chapter concludes with remarks regarding implications of this review for practice and research.

5.1 The outcome of this review

The objective of this review was to evaluate and compare the measurement properties of instruments that measure collaboration, specifically those which have been psychometrically tested and validated and are generalizable to various HCS. To achieve this, the review set out to identify studies on the measurement properties of instruments that measure collaboration, assess the studies for methodological quality and synthesize and present the results.

Through a process of database searching, report retrieval and exclusion of irrelevant papers it was possible to identify 21 studies encompassing 12 different instruments that produced validity data with samples of mixed participant types. It was decided to exclude studies measuring collaboration between only two participant types. Although the utility of such tools is useful in some settings, there is a need for tools that can assess collaboration in settings not defined by interprofessional collaboration only. However, some instruments included in this review were found to be applicable to a narrower definition of participant types and their use in some HCS may be invalid. For example the DOC questionnaire is valid for use with general physicians and specialists. This study was included in this review on the basis that specialists represent a diversity of

participant types. The study by Nuno-Solinis et al.[160] looked at collaboration between levels of healthcare using nurses and doctors as participants. Again, this study was included as the participant mix was adequately diverse as both nurses and doctors were diverse in their specialties. The OTAS checklist was designed for assessment of team behavior in the OR and much of the development and utilization of the OTAS relates to the OR setting. However, the study by O'Leary et al.[111] demonstrated the potential of the OTAS to be adapted to other HCSs. Even though the participant types in the OTAS studies relate mainly to OR staff, the interdisciplinary diversity was adequate to include these studies in this review.

The Index of Interdisciplinary Collaboration (IIC) and its modified versions represent important instruments in the measurement of collaboration because of strong evidence of validity and its expanding application in various care settings. However, it was decided to exclude the development report by Bronstein[27] for critical appraisal because the study sample was social workers only and not a significant mix of participant types. This study forms a significant component of the development of the IIC and the assessment of validity. Subsequent studies utilizing the modified IIC (MIIC)[148-151] were included in this review because they met the inclusion criteria for the sample. These studies refer to the Bronstein report for validation and contribute data useful in assessing interpretability of the IIC. Most of the studies using the modified versions of the MIIC,[148-151] referred to the factor analysis of Bronstein.[27] Only the validation study of the IITC-ESMH[152] performed an exploratory factor analysis which produced the same factor structure as Bronstein.[27]

The studies utilizing the MIIC utilized samples with a mix of participants that included non-healthcare professionals (chaplain, administration, social work). Similarly, the study of the CPAT tool included spiritual care and social workers.[87] All of the 21 studies sampled health or social care professionals predominantly and excluded patients or their families in the sample with two exceptions. The CPSET studies by Vanhaecht et al.[169]

included 12 patients in the pilot testing and patients contributed to prior focus groups and content validation. Although the study report indicated the participants in the structural validation of the beta version CPSET was represented by "51 patient groups", it was unclear as to whether any patients returned questionnaires, and in the report it appears that the 511 returned questionnaires was provided by healthcare professionals only. The IPS study by Korner and Wirtz[161] included patients in the measurement of internal participation by quantifying the patient's perception of the interprofessional team's collaboration.

Considering the growing awareness of including patient and families in collaborative healthcare[133-136] it would be prudent to develop and evaluate measurement tools with samples that better reflect these principles. It is suggested that including patients in the developmental processes, including content and face validity and the validity and reliability testing phase, will contribute to validation in a mixed sample that includes non-healthcare persons.

The COSMIN checklist proved its utility in assessing methodological quality in appraising the studies. To achieve this, an Excel spreadsheet was developed for the purpose of recording appraisal data from two independent appraisers (SJW and JJ). Discrepancies between appraisers were resolved by discussion and achieving consensus. Unresolved discrepancies were resolved by a third independent appraiser (SRM). The appraisal data was used to table the results as the quality of each instrument's measurement properties. Data extracted from the generalizability section of COSMIN was used to tabulate the characteristics of the included studies.

All studies appraised in this review performed a partial validation. No instrument had undergone completed evaluation of all validity and reliability properties. Content validity was addressed in all but one study[161] and the methodological quality across studies was fair to excellent. The methodological quality of structural validation and internal consistency varied between studies, however all primary development studies extracted a

factor structure. Test-retest, inter-rater and intra-rater reliability assessment, hypothesis testing and criterion validity was lacking in some studies. An instrument with good test-retest reliability can be depended upon to provide a measure of collaboration without introducing measurement error. More reliability testing is needed for the instruments reviewed in this study. Furthermore, as most instruments are founded on well established theories on collaboration, it would be desirable to conduct hypothesis testing as a component of construct validation. Also, the availability of instruments measuring collaboration and other aspects of team function allows for testing of convergent validity and therefore developers should extended validity tests to include these analyses.

No studies considered responsiveness of the instrument. Considering the utility of measuring collaboration in assessing teamwork, the ability of an instrument to detect changes in team dynamics is an important criterion. Longitudinal studies are needed to validate an instrument's capacity to assess changes in collaboration within the healthcare setting.

Two study reports mentioned measurement equivalence[85, 171] (alternatively measurement invariance or metric equivalence[85]) but no studies assessed for it. The problem of measurement equivalence is stated by Horn and McArdle;[102(p117)]

> *"The general question of invariance of measurement is one of whether or not, under different conditions of observing and studying phenomena, measurements yield measures of the same attributes".*

This means that when measurement is conducted to compare one group with another, without evidence of the presence or absence of measurement equivalence, it is not possible to interpret the finding of differences between individuals or groups.

The complexity of the healthcare setting, specifically relating to participant types, variation in workplace culture, types of healthcare practice, variation within and between professions and location of practice to mention a few, may impact on an individual's interpretation of a questionnaire's items. Therefore, different participant types as

individuals and as groups, may interpret items differently. Measurement equivalence should be a prerequisite test to hypothesis evaluation of group differences[175] and should be implemented in studies of measurement properties of instruments measuring collaboration especially where participant types are mixed.

All instruments in this review were developed using a Classical Test Theory (CTT) paradigm. The utility of Item Response Theory (IRT) for this area of collaboration measurement is still to be explored. The instruments were developed by addressing specific areas of collaboration including measuring the variance in; perception; people's beliefs, behaviors and attitudes; between different levels of care; multi-rater on target groups; relationships; the organization of care processes; and internal participation, reflecting a diversity of approaches to measuring collaboration. We posit that collaboration as a social phenomenon is universal. Regardless of the theoretical underpinnings adopted, or the unique research questions asked, or the uniqueness of the measurement instrument, the latent variables that comprise a measurement of collaboration should manifest coherently.

Using a multiple method approach (triangulation) the observation that perception of collaboration and actual collaborative acts differ[149] suggests the potential for using qualitative methods as a component of reliability and validity.[194] Observed differences may suggest the content of a questionnaire may not reflect the construct being measured. As posited by Denzin;[195(p82)]

> *"The use of multiple methods, or triangulation, reflects an attempt to secure an in-depth understanding of the phenomenon in question. Objective reality can never be captured. We only know a thing through its representations. Triangulation is not a tool or a strategy of validation but an alternative to validation. The combination of multiple methodological practices, empirical materials, perspectives, and observers in a single study is best understood as a strategy that adds rigor, breadth, complexity, richness, and depth to any inquiry".*

Though there is a diversity of theoretical under-pinning and real world applications, measurement of collaboration in its various forms, seems to have a cohesive underlying

structure as indicated by the synthesis performed with the latent variables in this review (see Table 4.3). This adds support to the various theories of collaboration and suggests the shared meaning of collaboration as a social phenomenon.

The inherent complexity of collaboration as a phenomenon requires a research approach to measurement instrument development that incorporates complexity, so that the measurement instrument's validity and reliably captures a metric of correlation value.

Many instruments for the measurement of collaboration within HCS exist. However, the quality of each instrument varies; instruments are designed for specific populations and purposes and are validated in various settings. Selecting an instrument requires careful consideration of the qualities of each. Therefore, referring to systematic reviews of measurement properties of instruments may be helpful to clinicians or researchers in instrument selection. This review identified, appraised and presented the measurement properties of 12 validated instruments and described the characteristics of the sample for which the instruments were validated. Furthermore, a narrative synthesis of the factor structure of these instruments produced nine factor categories that reflect the theoretical constructs of the measurement of collaboration within HCS.

5.2 Measuring complexity

Collaboration is an essential trait for human survival and possibly the most complex of all social behaviors. As stated by D'Amour et al., "...[o]ur working lives are set in collective environments with constant interactions with others".[146(p116)] The HCS is an environment in which the interactions between participants form the basis for caring and curing. Collaboration is an effective interaction especially if caring and curing are the desired outcomes. If the coming-together of stakeholders to solve a problem by forming a collaborative team is an advantage, how do we explain that advantage?

It is posited that through collaborative effort the issue of *complexity* may be addressed. Myra Wilson suggests "...complexity is a fact of life in healthcare...".[4(p19)]

Through collaboration we are able to bring together skills and knowledge that allows us to address the complexity of healthcare and provide the best patient care.[18(p715)] To understand complexity it is helpful to refer to complexity theory.

Complexity theory evolved initially in the 1980s and offered a complementary view of phenomena beyond the Newtonian reductionist view of machine-like systems and nested subsystems.[4] Complexity incorporates a view that considers the interconnectedness of elements and the importance of the environment in which the elements exist. This is known as a Complex Adaptive System(CAS).[3] A CAS is defined as a collection of individual agents, who act freely in ways that are not always predictable and whose actions have an effect on other agents within the system.[3] An example of a complex network is the immune system[196] and so too the human body.[197] Furthermore, any collection of people is a CAS including healthcare teams.[3] Collaboration within healthcare settings reflects this type of complexity. Additionally, a contribution to this complexity is the biopsychosocial nature of health problems.

5.3 The biopsychosocial model

The term biopsychosocial is "...a term meaning to consider a persons' biological, psychological and social makeup as a way of viewing the human condition as a continuum of connected and nested hierarchies".[1(px)] Since George Engel published his first paper titled 'The Need for a New Medical Model: A Challenge for Biomedicine' in 1977[198] there has been a growing interest in and adoption of the biopsychosocial approach within the healthcare setting. A current PubMed search (6/10/2014) of titles and abstracts using the search term 'biopsychosocial' returns 3463 papers, of which 962 are review papers, 129 clinical trials and 64 randomized controlled trials. One paper was published in 1974 compared to 270 papers published in 2013. This indicates a growing interest and importance of the biopsychosocial model in healthcare.

Adopting a biopsychosocial model may address the inherent complexity of caring

for any one patient with a unique health problem. For example, a patient suffering with multiple recurrent bacterial infections might be treated with antibiotics to kill the microbes. If however, the patient had been suffering with high levels of persistent stress due to impoverishment, poor nutrition and living in unsanitary conditions, is the problem bacterium? The bacterial infection is a symptom of the disease (bio), chronic stress reduces immunity (psycho) and impoverishment (social) leads to poor nutrition and substandard living conditions, which are all inter-related. Unless the patient is treated in a wholistic (biopsychosocial) way, health may not be re-established.

Because the practice of modern healthcare has become highly specialized, it is not possible for one practitioner to address all layers of the patient's biopsychosocial health. For complex health problems like chronic conditions, effective healthcare must be multidisciplinary.

The biopsychosocial concept underpins the benefit of collaboration in healthcare settings; the coalescing of various healthcare and non-healthcare approaches to caring in a coordinated, cooperative and communicative approach to human care to address the complexity of the biopsychosocial whole.

5.4 The value of factor analysis

Studies evaluating structural validity of instruments using factor analysis produce data in the form of factor structures. This essentially represents a reduction of complex variability data down to a parsimonious description of latent variables and is referred to as simple structure. In other words, factor analysis accounts for the largest amount of shared variance of the smallest number of latent variables.

Multiple factor analysis was described by Thurstone in 1931[199] and elaborated in 1947.[200] Despite the longterm, widespread use and trust in this method of correlation analysis, it is not without criticism.[201-204] Ertel suggests that "...Thurstone's parsimony needs reconsideration"[201(p196)] and;

> *"The mathematical simplicity of simple structure, destroying factorial combinations, is imposed, tacitly and blind, on seemingly solitary observational entities ("variables") while the underlying components of these entities are entirely ignored. Simple structure rotation forces variables into clusters while the sources of clustering remain obscure. Empirical research demands an unveiling of relations among functional components, but this demand is obstructed by Thurstone's doubtful methodical decision".*[201(p196)]

Ertel's solution is to rotate data using his Varimin method in order to minimize variance and hence to capture the complexity of the underlying data.

This discussion challenges the methods of mathematical treatment of data acquired using measurement and the reduction of this data to simple structure that assumes factors are discrete entities, which is doubtful. The discovery and measurement of factors in instrument development obliterates the relationships between factors. Regardless of these issues, factor analysis remains a common treatment in the development of psychometric instruments and models of complex phenomena.

In this current study the social phenomenon of collaboration within healthcare settings was considered from a measurement problem perspective. The studies included produced validity data that revealed useful information regarding instrument selection and development. One part of this data was structural validity data in the form of factor structures for each instrument. One instrument, the OTAS was not assessed for structural validity and hence no factor structure was available. One instrument, the IPS, produced a one factor solution.

It was decided to explore the data with a synthesis of the results of each instrument's factor analysis. In other words, the intention was to create a summary of the simple structures from each study. For this purpose a process was used to perform a narrative synthesis. By examining the factor descriptors (a text heading) of each instrument's factor structure, and by creating a common table, each factor descriptor was allocated to a column of the table. (see Table 4.3). The process involved grouping similar

factor descriptors for each instrument under the same columns.

To overcome the difficulty of allocating some factor descriptors, a closer examination of the observable variables (the related question(s) grouped under a factor descriptor) was undertaken to understand the meaning of the factor descriptor. To be inclusive, the OTAS was allocated to columns using the model's theoretical categories (based on an input-process-output model)[205] for grouping the checklist questions for collaboration; leadership, monitoring, communication, coordination and cooperation. Because the IPS produced a single factor structure, the six observable variables were used for allocation to the table.

Narrative synthesis of simple structure data across multiple studies is possibly a novel treatment of structural validity data. A question remains about the utility, method and validity of this approach. Some future research is needed to determine if this type of data treatment is helpful in refining or confirming theoretical models.

5.5 Implications for practice

Systematic reviews of measurement properties provide practical guidelines for the selection of instruments as well as the ongoing development and validation of measurement tools. For practice settings that have a focus on an integrative medicine approach and where managers evaluate their teams for efficacy, it may not be enough to assume teams collaborate without also obtaining an empirical measurement. Furthermore, evaluating collaboration provides important information on the strengths and limitations of different HCS and the opportunities for continuous improvement via any remedial actions initiated.

The tools evaluated in this review have partial evidence of validity and further validation studies are recommended. However all of the instruments have potential to be considered for use in clinical settings, for research measurements and for ongoing validation studies. It is recommended that readers evaluate the characteristics of individual

studies presented in this review and align the sample and setting with their own circumstances.

- Tool selection should also align with the purpose of measurement. For example, if the purpose of the instrument is to assess team collaboration the CPAT, CPSET, HTVI, IITC-ESMH, OTAS or the IIC/MIIC might be chosen.
- If it is desirable to assess team collaboration between two levels of care, the IPC-DLC may be useful.
- If measuring belief, behavior and attitudes towards collaboration is an objective, the ISVS and the DOC are appropriate.
- Based on assessing the collaborative relationships the AITCS may be chosen and the MGMS used for multiple groups rating collaboration with each other.
- To assess the interprofessional team for internal participation from the team members' and from the patient's perception, the IPS is specific for this purpose.

5.6 Implications for research

There is a need for further validation studies for the instruments presented in this current study. Test-retest and interrater reliability would be a valuable contribution to instrument validation. Responsiveness validation is needed, as instruments that measure collaboration in HCS are valuable when assessing team function, especially when assessing improvement due to any remedial processes.

The instruments assessed in this current study were based on the measurement of perception or observable behaviors and were developed using latent variable methodologies. A question remains as to whether a test could be developed using Item

Response Theory methodologies to measure the collaborative trait level for an individual? By measuring an individuals' trait level one is evaluating an element within a complex adaptive system.

A result of the proliferation of research on collaboration measurement is the accumulation of a data-bank of items that could be used to develop item response models for collaboration measurement. The translation of IRT models into Computer Adaptive Tests (CAT) could be an effective way of measuring trait levels in individuals within a variety of HCS. The use of CAT has the advantage of determining trait levels with an individual's response to a minimal number of questions facilitated by computer input. This may provide a practical method of measuring collaboration in HCS without creating additional burden to collaborators or administrators and assisting researchers in data collection and model refinement.

Important to practice is the development of a tool that can be used to measure collaboration across both interprofessional teams and non-professionals such as with patients and families. Measurement equivalence is an important consideration for future instrument development and validation. Further development of the COSMIN tool should include critical appraisal for measurement equivalence.

It is proposed that, in addition to extending tool development using IRT modelling or using measurement equivalence as a component of validation, that a multiple method (triangulation) approach to instrument validation be considered. The use of observational measurements of collaborative acts (e.g. the OTAS instrument) could be used to compare against measurements of perception of collaboration.

Finally, it is recommended that researchers identify and report the practice style of the collaborative team, because practice style is an independent variable when conducting research within HCS. Further research could develop a comprehensive model of practice style and research instruments to assist researchers classify healthcare teams.

Conclusion

This study set out to evaluate the state of measurement of collaboration in HCS using a systematic review of measurement properties of instruments. The result was the discovery of 21 studies of 12 instruments that measure collaboration in HCS with a complex mix of participant types.

Though all studies produced useful validity data and all 12 instruments show potential for utility in a range of HCS, the difficulty and accuracy of measuring collaboration relates to the complexity of the HCS. Detailed consideration of complexity is necessary if measurement of collaboration in HCS is to provide accurate and useful data. Whether traditional measurement models such as CTT are adequate in capturing the effect of complexity when measuring phenomena such as collaboration was challenged in this study, however the enduring tradition and trust in this theory dominates in healthcare research. There is considerable scope to look at other measurement models such as IRT and using multi-method measurements. Furthermore, even though models of practice style have been proposed, current research pays little attention to classifying teams despite this being an important independent variable.

Considering the importance of collaboration in healthcare to improve patient safety, patient health outcomes and practitioner satisfaction, valid and practical methods of measurement are critically needed. This study has presented research that goes some way to identifying the current state of measurement science in this research domain.

References

1. Walters SJ. Fibromyalgia and stress; addressing the biopsychosocial whole. Armidale, New South Wales: University of New England; 2005.
2. Wood DJ, Gray B. Toward a Comprehensive Theory of Collaboration. J Appl Behav Sci. 1991 June 1, 1991;27(2):139-62.
3. Plsek PE, Greenhalgh T. The challenge of complexity in health care. BMJ. [Journal Article]. 2001;323(7313):625-8.
4. Wilson M. Complexity Theory. Whitireia Nursing Journal. 2009(16):18-24.
5. Pearson A, Wiechula R, Court A, Lockwood C. The JBI model of evidence-based healthcare. Int J Evid Based Healthc. 2005 Sep;3(8):207-15.
6. Athanasiou T, Darzi A. Evidence Synthesis in Healthcare In: Athanasiou T, Darzi A, editors. A Practical Handbook for Clinicians. London: Springer; 2011. p. 1-46.
7. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010 Jul;63(7):737-45.
8. Saso S, Panesar P, Weiming S, Athanasiou T. Systematic review and meta-analysis in clinical practice In: Athanasiou T, Darzi A, editors. Evidence synthesis in healthcare. London: Springer; 2011. p. 67-114.
9. Messick S. Validity. In: Linn RL, editor. Educational measurement. 3rd. ed. New York: American Council on Education/Macmillan; 1989. p. 13–103.
10. America IoMCoQoHCi. To Err Is Human : Building a Safer Health System. Kohn LT, Corrigan J, Donaldson MS, Institute of M, editors: National Academies Press; 2000.
11. America IoMCoQoHCi. Crossing the Quality Chasm : A New Health System for the 21st Century: National Academies Press; 2001.
12. Australian Government Department of Health. Chronic Disease. Canberra: Australian Government; 2012. Available from: http://www.sti.health.gov.au/internet/main/publishing.nsf/Content/chronic
13. Leininger M. The phenomenon of caring: Caring: The essence and central focus of nursing. Nurs Res Rep. 1977;12(1):2,14.
14. Frampton SB, Charmel PA, Guastello S. The Putting Patients First Field Guide Global Lessons in Designing and Implementing Patient-Centered Care. Hoboken: Wiley; 2013.
15. Gaboury I, Lapierre LM, Boon H, Moher D. Interprofessional collaboration within integrative healthcare clinics through the lens of the relationship-centered care model. J Interprof Care. 2011 Mar;25(2):124-30.
16. Gaboury I, Bujold M, Boon H, Moher D. Interprofessional collaboration within Canadian integrative healthcare clinics: Key components. Soc Sci Med. 2009;69(5):707-15.
17. Gaboury I, Boon H, Verhoef M, Bujold M, Lapierre LM, Moher D. Practitioners' validation of framework of team-oriented practice models in integrative health care: a mixed methods study. BMC Health Serv Res. 2010;10:289.

18. Boon HS, Mior SA, Barnsley J, Ashbury FD, Haig R. The difference between integration and collaboration in patient care: results from key informant interviews working in multiprofessional health care teams. J Manipulative Physiol Ther. 2009 Nov-Dec;32(9):715-22.
19. Boon HS, Kachan N. Integrative medicine: a tale of two clinics. BMC Complement Altern Med. 2008;8:32.
20. Boon H, Verhoef M, O'Hara D, Findlay B, Majid N. Integrative healthcare: arriving at a working definition. Altern Ther Health Med. 2004 Sep-Oct;10(5):48-56.
21. Boon H, Verhoef M, O'Hara D, Findlay B. From parallel practice to integrative health care: a conceptual framework. BMC Health Serv Res. 2004 Jul 1;4(1):15.
22. Chamberlain-Salaun J, Mills J, Usher K. Terminology used to describe health care teams: an integrative review of the literature. J Multidiscip Healthc. 2013;6:65-74.
23. Feldman LS, Costa LL, Feroli ER, Jr., Nelson T, Poe SS, Frick KD, et al. Nurse-pharmacist collaboration on medication reconciliation prevents potential harm. J Hosp Med. 2012 May-Jun;7(5):396-401.
24. Zwarenstein M, Goldman J, Reeves S. Interprofessional collaboration: effects of practice-based interventions on professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2009;3(CD000072):1-31.
25. Chung B, Dopheide JA, Gregerson P. Psychiatric pharmacist and primary care collaboration at a skid-row safety-net clinic. J Natl Med Assoc. 2011 Jul;103(7):567-74.
26. Sullivan TJ. Collaboration: A Health Care Imperative. New York, NY: McGraw-Hill; 1998.
27. Bronstein LR. Instrument development. Index of interdisciplinary collaboration. Soc Work Res. 2002;26(2):113-22.
28. D'Amour D, Ferrada-Videla M, San Martin Rodriguez L, Beaulieu M-D. The conceptual basis for interprofessional collaboration: Core concepts and theoretical frameworks. J Interprof Care. 2005;19(S1):116-31.
29. San Martín-Rodríguez L, Beaulieu M, D'Amour D, Ferrada-Videla M. The determinants of successful collaboration: a review of theoretical and empirical studies. J Interprof Care. 2005;19:132-47.
30. Dictionary OE. "conceit, n.": Oxford University Press.
31. The Oxford Dictionary of Sports Science & Medicine 3ed: Oxford University Press; 2006.
32. Dictionary OE. "characteristic, n. and adj.": Oxford University Press.
33. Dictionary OE. "attribute, n.": Oxford University Press.
34. Orchard CA, King GA, Khalili H, Bezzina MB. Assessment of Interprofessional Team Collaboration Scale (AITCS): development and testing of the instrument. J Contin Educ Health Prof. 2012 Winter;32(1):58-67.
35. Bruner C. Thinking collaboratively: ten questions and answers to help policy makers improve children's services Washington, D.C.: Education and Human Services Consortium 1991. p. 27.
36. Evans JA. The role of the nurse manager in creating an environment for collaborative practice. Holist Nurs Pract. 1994;8(3):22-31.

37. Nugus P, Greenfield D, Travaglia J, Westbrook J, Braithwaite J. How and where clinicians exercise power: interprofessional relations in health care. Soc Sci Med. 2010 Sep;71(5):898-909.
38. Chung VC, Ma PH, Hong LC, Griffiths SM. Organizational determinants of interprofessional collaboration in integrative health care: systematic review of qualitative studies. PLoS One. 2012;7(11):e50022.
39. MacNaughton K, Chreim S, Bourgeault IL. Role construction and boundaries in interprofessional primary health care teams: a qualitative study. BMC Health Serv Res. 2013;13:486.
40. Hoppe MH. Active Listening : Improve Your Ability to Listen and Lead [Internet]. Greensboro, NC, USA: Center for Creative Leadership; 2006. Available from: http://site.ebrary.com/lib/adelaide/docDetail.action?docID=10193835
41. Eriksson K. Evidence: To See or Not to See. Nursing Science Quarterly. 2010 October 1, 2010;23(4):275-9.
42. Edwards M. Keywords in the history of medicine: Evidence. Lancet. [Article]. 2004;363(9421):1657-.
43. Guba E, Lincoln Y. Epistemological and methodological bases of naturalistic inquiry. ECTJ. 1982;30(4):233-52.
44. Staller KM. Qualitative Research. In: Salkind NJ, editor. Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications, Inc.; 2010. p. 1159-64.
45. Kraska M. Quantitative Research. In: Salkind NJ, editor. Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications, Inc.; 2010. p. 1167-72.
46. CEBM. Levels of Evidence. University of Oxford; 2009. Available from: http://www.cebm.net/index.aspx?o=1025
47. Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. BMJ. [Journal Article]. 2001;323(7308):334-6.
48. JBI. Reviewer's Manual 2014.
49. Egger M, Smith GD. Meta-Analysis. Potentials and promise. BMJ. 1997 Nov 22;315(7119):1371-4.
50. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statist Med. 2002;21:1539-58.
51. Traub RE. Classical Test Theory in Historical Perspective. Educational Measurement: Issues and Practice. 1997;16(4):8-14.
52. Viswanathan M. Measurement Error and Research Design. Thousand Oaks, California: Sage Publications; 2005.
53. Webb NM, Shavelson RJ, Haertel EH. 4 Reliability Coefficients and Generalizability Theory. In: Rao CR, Sinharay S, editors. Handbook of Statistics: Elsevier; 2006. p. 81-124.
54. Graham JM. Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability: What They Are and How to Use Them. Educ Psychol Meas. 2006 December 1, 2006;66(6):930-44.
55. Rogers JH. Item Response Theory. In: Salkind NJ, editor. Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications, Inc.; 2010. p. 646-52.

56. Harvey RJ, Hammer AL. Item Response Theory. Couns Psychol. 1999 May 1, 1999;27(3):353-83.
57. Bock RD. A Brief History of Item Theory Response. Educational Measurement: Issues and Practice. 1997;16(4):21-33.
58. Thissen D, Steinberg L. A taxonomy of item response models. Psychometrika. 1986;51(4):567-77.
59. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In: Lord FM, Novick MR, editors. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley; 1968.
60. Lord FM, Novick MR. Statistical theories of mental test scores. Reading , MA: Addison-Wesley; 1968.
61. van der Linden WJ. Item Response Theory. In: McGaw PPB, editor. International Encyclopedia of Education (Third Edition). Oxford: Elsevier; 2010. p. 81-8.
62. Wanders RB, Wardenaar KJ, Kessler RC, Penninx BW, Meijer RR, de Jonge P. Differential reporting of depressive symptoms across distinct clinical subpopulations: What DIFference does it make? J Psychosom Res. 2015;78(2):130-6.
63. Wahl I, Lowe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, et al. Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. J Clin Epidemiol. 2014 Jan;67(1):73-86.
64. Victorson DE, Choi S, Judson MA, Cella D. Development and testing of item response theory-based item banks and short forms for eye, skin and lung problems in sarcoidosis. Qual Life Res. 2014 May;23(4):1301-13.
65. Victorson D, Cavazos JE, Holmes GL, Reder AT, Wojna V, Nowinski C, et al. Validity of the Neurology Quality-of-Life (Neuro-QoL) measurement system in adult epilepsy. Epilepsy Behav. 2014 Feb;31:77-84.
66. Sun X, Allison C, Auyeung B, Matthews FE, Norton S, Baron-Cohen S, et al. Psychometric properties of the Mandarin version of the Childhood Autism Spectrum Test (CAST): an exploratory study. J Autism Dev Disord. 2014 Jul;44(7):1565-76.
67. Sharp C, Steinberg L, Temple J, Newlin E. An 11-item measure to assess borderline traits in adolescents: refinement of the BPFSC using IRT. Personal Disord. 2014 Jan;5(1):70-8.
68. Schalet BD, Cook KF, Choi SW, Cella D. Establishing a common metric for self-reported anxiety: linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. J Anxiety Disord. 2014 Jan;28(1):88-96.
69. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE, Jr. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol. 2014 May;67(5):516-26.
70. Ravens-Sieberer U, Herdman M, Devine J, Otto C, Bullinger M, Rose M, et al. The European KIDSCREEN approach to measure quality of life and well-being in children: development, current application, and future advances. Qual Life Res. 2014 Apr;23(3):791-803.
71. Pardasaney PK, Ni P, Slavin MD, Latham NK, Wagenaar RC, Bean J, et al. Computer-adaptive balance testing improves discrimination between community-dwelling elderly fallers and nonfallers. Arch Phys Med Rehabil. 2014 Jul;95(7):1320-7 e1.

72. McGrory S, Doherty JM, Austin EJ, Starr JM, Shenkin SD. Item response theory analysis of cognitive tests in people with dementia: a systematic review. BMC Psychiatry. 2014;14:47.

73. Hassani L, Dehdari T, Hajizadeh E, Shojaeizadeh D, Abedini M, Nedjat S. Development of an instrument based on the protection motivation theory to measure factors influencing women's intention to first pap test practice. Asian Pac J Cancer Prev. 2014;15(3):1227-32.

74. Guillen V, Santos B, Munoz P, Fernandez de Corres B, Fernandez E, Perez I, et al. Toronto alexithymia scale for patients with eating disorder: of performance using the non-parametric item response theory. Compr Psychiatry. 2014 Jul;55(5):1285-91.

75. Godefroy O, Gibbons L, Diouf M, Nyenhuis D, Roussel M, Black S, et al. Validation of an integrated method for determining cognitive ability: Implications for routine assessments and clinical trials. Cortex. 2014 May;54:51-62.

76. Gerrard P, Zafonte R, Giacino JT. Coma recovery scale-revised: evidentiary support for hierarchical grading of level of consciousness. Arch Phys Med Rehabil. 2014 Dec;95(12):2335-41.

77. Fries JF, Witter J, Rose M, Cella D, Khanna D, Morgan-DeWitt E. Item response theory, computerized adaptive testing, and PROMIS: assessment of physical function. J Rheumatol. 2014 Jan;41(1):153-8.

78. Fries JF, Lingala B, Siemons L, Glas CA, Cella D, Hussain YN, et al. Extending the floor and the ceiling for assessment of physical function. Arthritis Rheumatol. 2014 May;66(5):1378-87.

79. Farin E, Nagl M, Gramm L, Heyduck K, Glattacker M. Development and evaluation of the PI-G: a three-scale measure based on the German translation of the PROMIS (R) pain interference item bank. Qual Life Res. 2014 May;23(4):1255-65.

80. Dmitrieva NO, Fyffe D, Mukherjee S, Fieo R, Zahodne LB, Hamilton J, et al. Demographic characteristics do not decrease the utility of depressive symptoms assessments: examining the practical impact of item bias in four heterogeneous samples of older adults. Int J Geriatr Psychiatry. 2015 Jan;30(1):88-96.

81. Cheville AL, Wang C, Ni P, Jette AM, Basford JR. Age, sex, and symptom intensity influence test taking parameters on functional patient-reported outcomes. Am J Phys Med Rehabil. 2014 Nov;93(11):931-7.

82. Cheville AL, Basford JR, Dos Santos K, Kroenke K. Symptom burden and comorbidities impact the consistency of responses on patient-reported functional outcomes. Arch Phys Med Rehabil. 2014 Jan;95(1):79-86.

83. Carle AC, Jean-Pierre P, Winters P, Valverde P, Wells K, Simon M, et al. Psychometric evaluation of the patient satisfaction with logistical aspects of navigation (PSN-L) scale using item response theory. Med Care. 2014 Apr;52(4):354-61.

84. Hull L, Arora S, Kassab E, Kneebone R, Sevdalis N. Observational Teamwork Assessment for Surgery: Content Validation and Tool Refinement. J Am Coll Surg. 2011;212(2):234-43.e5.

85. Passauer-Baierl S, Hull L, Miskovic D, Russ S, Sevdalis N, Weigl M. Re-validating the observational teamwork assessment for surgery tool (OTAS-D):

cultural adaptation, refinement, and psychometric evaluation. World J Surg. 2014 Feb;38(2):305-13.

86. Wagner R, Thatcher Kantor P, Piasta S. Latent Variable. In: Salkind NJ, editor. Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications, Inc.; 2010. p. 697-9.

87. Schroder C, Medves J, Paterson M, Byrnes V, Chapman C, O'Riordan A, et al. Development and pilot testing of the collaborative practice assessment tool. J Interprof Care. 2011 May;25(3):189-95.

88. Fritchoff Davis M. Method Variance. In: Salkind NJ, editor. Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications, Inc.; 2010. p. 802-5.

89. Rencher AC, Christensen WF. Exploratory factor analysis. Wiley Series in Probability and Statistics : Methods of Multivariate Analysis (3rd Edition). Somerset, NJ, USA: John Wiley & Sons; 2012.

90. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;2(4):433-59.

91. Coleman JSM. Principal Components Analysis. In: Salkind NJ, editor. Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications, Inc.; 2010. p. 1098-103.

92. Hayasbi K, Yuan K. Exploratory Factor Analysis. In: Salkind NJ, editor. Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications, Inc.; 2010. p. 459-66.

93. Rencher AC, Christensen WF. Confirmatory factor analysis. Wiley Series in Probability and Statistics : Methods of Multivariate Analysis (3rd Edition). Somerset, NJ, USA: John Wiley & Sons; 2012. p. 479-500.

94. Shafer AB. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. J Clin Psychol. 2006 Jan;62(1):123-46.

95. Shafer A. Meta-analysis of the brief psychiatric rating scale factor structure. Psychol Assess. 2005 Sep;17(3):324-35.

96. Kim G, Decoster J, Huang CH, Chiriboga DA. Race/ethnicity and the factor structure of the Center for Epidemiologic Studies Depression Scale: a meta-analysis. Cultur Divers Ethnic Minor Psychol. 2011 Oct;17(4):381-96.

97. Kim G, DeCoster J, Huang CH, Bryant AN. A meta-analysis of the factor structure of the Geriatric Depression Scale (GDS): the effects of language. Int Psychogeriatr. 2013 Jan;25(1):71-81.

98. Frick PJ, Lahey BB, Loeber R, Tannenbaum L, Van Horn Y, Christ MAG, et al. Oppositional defiant disorder and conduct disorder: A meta-analytic review of factor analyses and cross-validation in a clinic sample. Clin Psychol Rev. 1993;13(4):319-40.

99. Sevdalis N, Lyons M, Healey AN, Undre S, Darzi A, Vincent CA. Observational teamwork assessment for surgery: construct validation with expert versus novice raters. Ann Surg. 2009 Jun;249(6):1047-51.

100. Sperber AD. Translation and validation of study instruments for cross-cultural research. Gastroenterol. 2004 Jan;126(1 Suppl 1):S124-8.

101. Pena ED. Lost in translation: methodological considerations in cross-cultural research. Child Dev. 2007 Jul-Aug;78(4):1255-64.

102. Horn JL, McArdle JJ. A practical and theoretical guide to measurement invariance in aging research. Exp Aging Res. 1992 Autumn-Winter;18(3-4):117-44.
103. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003 Feb;80(1):99-103.
104. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297-334.
105. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53-5.
106. Viera AJ, Garrett JM. Understanding Interobserver Agreement: kappa Family Medicine. 2005;37(5):360-3.
107. Multon KD. Interrater Reliability. In: Salkind NJ, editor. Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications, Inc.; 2010. p. 627-9.
108. Howell DC. Intraclass Correlation. In: Salkind NJ, editor. Encyclopedia of Research Design Thousand Oaks, CA: SAGE Publications, Inc.; 2010. p. 637-42.
109. Undre S, Sevdalis N, Healey AN, Darzi A, Vincent CA. Observational teamwork assessment for surgery (OTAS): refinement and application in urological surgery. World J Surg. 2007 Jul;31(7):1373-81.
110. Russ S, Hull L, Rout S, Vincent C, Darzi A, Sevdalis N. Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. Ann Surg. 2012 Apr;255(4):804-9.
111. O'Leary KJ, Boudreau YN, Creden AJ, Slade ME, Williams MV. Assessment of teamwork during structured interdisciplinary rounds on medical units. J Hosp Med. 2012 Nov-Dec;7(9):679-83.
112. Leighton JP. External Validity. In: Salkind NJ, editor. Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications, Inc.; 2010. p. 467-71.
113. Tebes JK. External validity and scientific psychology. Am Psychol. 2000;55(12):1508-9.
114. Cook TD, Campbell DT. Quasi-experimentation: design & analysis issues for field settings. Campbell DT, editor. Boston: Houghton Mifflin; 1979.
115. The Joanna Briggs Institute. Joanna Briggs Institute Reviewers' Manual: 2014 edition: Joanna Briggs Institute; 2014.
116. Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res. 2009 Apr;18(3):313-33.
117. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. BMC Med Res Methodol. 2006;6:2.
118. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. [Article]. 2010;19(4):539-49.

119. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007 Jan;60(1):34-42.
120. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol. 2010;10:22.
121. Terwee C, Mokkink L, Knol D, Ostelo RJG, Bouter L, Vet HW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012;21(4):651-7.
122. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res. 2012 May;21(4):659-70.
123. Reimers AK, Mess F, Bucksch J, Jekauc D, Woll A. Systematic review on measurement properties of questionnaires assessing the neighbourhood environment in the context of youth physical activity behaviour. BMC Public Health. 2013;13:461.
124. Paalman CH, Terwee CB, Jansma EP, Jansen LM. Instruments measuring externalizing mental health problems in immigrant ethnic minority youths: a systematic review of measurement properties. PLoS One. 2013;8(5):e63109.
125. Oftedal S, Bell KL, Mitchell LE, Davies PS, Ware RS, Boyd RN. A systematic review of the clinimetric properties of habitual physical activity measures in young children with a motor disability. Int J Pediatr. 2012;2012:976425.
126. van Tulder M, Furlan A, Bombardier C, Bouter L. Updated method guidelines for systematic reviews in the cochrane collaboration back review group. Spine (Phila Pa 1976). 2003 Jun 15;28(12):1290-9.
127. Furlan AD, Pennick V, Bombardier C, van Tulder M. 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. Spine (Phila Pa 1976). 2009 Aug 15;34(18):1929-41.
128. Schellingerhout JM, Heymans MW, Verhagen AP, de Vet HC, Koes BW, Terwee CB. Measurement properties of translated versions of neck-specific questionnaires: a systematic review. BMC Med Res Methodol. 2011;11:87.
129. Walters SJ, Roberston-Malt S, Stern C. The measurement of collaboration within healthcare settings: a systematic review protocol of measurement properties of instruments. . The JBI Database of Systematic Reviews and Implementation Reports 2015 13(7):24-43.
130. Reeves S. Interprofessional teamwork for health and social care [Internet]. Chichester, West Sussex: Blackwell Publishing; 2010. Available from: http://site.ebrary.com/lib/adelaide/Top?id=10395577
131. Sicotte C, D'Amour D, Moreault M-P. Interdisciplinary collaboration within Quebec community health care centres. Soc Sci Med. 2002;55(6):991-1003.
132. Nathanson BH, Henneman EA, Blonaisz ER, Doubleday ND, Lusardi P, Jodka PG. How much teamwork exists between nurses and junior doctors in the intensive care unit? J Adv Nurs. 2011;67(8):1817-23.
133. Kon AA. The shared decision-making continuum. JAMA. 2010;304(8):903-4.

134. Howe A. Can the patient be on our team? An operational approach to patient involvement in interprofessional approaches to safe care. J Interprof Care. 2006;20(5):527-34.
135. Robertson S, Pryde K, Evans K. Patient involvement in quality improvement: is it time we let children, young people and families take the lead? Arch Dis Child Educ Pract Ed. 2014 Feb;99(1):23-7.
136. WHO. Framework for action on interprofessional education and collaborative practice: World Health Organisation; 2010. Report No.: WHO/HRH/HPN/10.3.
137. Gray B. Collaborating; finding common ground for multiparty problems. San Francisco: Jossey-Bass; 1989.
138. Zwarenstein M, Reeves S. What's so great about collaboration?: we need more evidence and less rhetoric. BMJ. 2000;320(7241):1022.
139. Vera M, Perez-Pedrogo C, Huertas SE, Reyes-Rabanillo ML, Juarbe D, Huertas A, et al. Collaborative care for depressed patients with chronic medical conditions: a randomized trial in Puerto Rico. Psychiatr Serv. 2010 Feb;61(2):144-50.
140. Van Leeuwen Williams E, Unutzer J, Lee S, Noel PH. Collaborative depression care for the old-old: findings from the IMPACT trial. Am J Geriatr Psychiatry. 2009 Dec;17(12):1040-9.
141. van der Voort TY, van Meijel B, Goossens PJ, Renes J, Beekman AT, Kupka RW. Collaborative care for patients with bipolar disorder: a randomised controlled trial. BMC Psychiatry. 2011;11:133.
142. Rollman BL, Belnap BH. The Bypassing the Blues trial: collaborative care for post-CABG depression and implications for future research. Cleve Clin J Med. 2011 Aug;78 Suppl 1:S4-12.
143. Muntingh AD, van der Feltz-Cornelis CM, van Marwijk HW, Spinhoven P, Assendelft WJ, de Waal MW, et al. Collaborative stepped care for anxiety disorders in primary care: aims and design of a randomized controlled trial. BMC Health Serv Res. 2009;9:159.
144. Johnson JA, Al Sayah F, Wozniak L, Rees S, Soprovich A, Chik CL, et al. Controlled trial of a collaborative primary care team model for patients with diabetes and depression: rationale and design for a comprehensive evaluation. BMC Health Serv Res. 2012;12:258.
145. Wilson SF, Marks R, Collins N, Warner B, Frick L. Benefits of multidisciplinary case conferencing using audiovisual compared with telephone communication: a randomized controlled trial. J Telemed Telecare. 2004;10(6):351-4.
146. D'Amour D, Ferrada-Videla M, San Martin Rodriguez L, Beaulieu M. The conceptual basis for interprofessional collaboration: core concepts and theoretical frameworks. J Interprof Care. 2005;19(S1):116-31.
147. Thannhauser J, Russell-Mayhew S, Scott C. Measures of interprofessional education and collaboration. J Interprof Care. 2010;24(4):336-49.
148. Wittenberg-Lyles EM, Parker Oliver D. The power of interdisciplinary collaboration in hospice. Prog Palliat Care. 2007;15(1):6-12.
149. Wittenberg-Lyles E, Parker Oliver D, Demiris G, Regehr K. Interdisciplinary collaboration in hospice team meetings. J Interprof Care. 2010;24(3):264-73.

150. Oliver DP, Wittenberg-Lyles EM, Day M. Measuring interdisciplinary perceptions of collaboration on hospice teams. Am J Hosp Palliat Care. [Empirical Study; Quantitative Study]. 2007 Feb-Mar;24(1):49-53.
151. Oliver DP, Wittenberg-Lyles EM, Day M. Variances in perceptions of interdisciplinary collaboration by hospice staff. J Palliat Care. 2006 Winter;22(4):275-80.
152. Mellin EA, Bronstein L, Anderson-Butcher D, Amorose AJ, Ball A, Green J. Measuring interprofessional team collaboration in expanded school mental health: Model refinement and scale development. J Interprof Care. 2010 Sep;24(5):514-23.
153. King G, Shaw L, Orchard CA, Miller S. The interprofessional socialization and valuing scale: a tool for evaluating the shift toward collaborative care approaches in health care settings. Work. 2010;35(1):77-85.
154. Seys D, Deneckere S, Sermeus W, Van Gerven E, Panella M, Bruyneel L, et al. The Care Process Self-Evaluation Tool: a valid and reliable instrument for measuring care process organization of health care teams. BMC Health Serv Res. 2013;13:325.
155. Berendsen AJ, Benneker WH, Groenier KH, Schuling J, Grol RP, Meyboom-de Jong B. DOC questionnaire: measuring how GPs and medical specialists rate collaboration. Int J Health Care Qual Assur. 2010;23(5):516-26.
156. Dougherty M, Larson E. A review of instruments measuring nurse-physician collaboration. J Nurs Adm. 2005;35(5):244-53.
157. Dedrick RF, Greenbaum PE. Multilevel confirmatory factor analysis of a scale measuring interagency collaboration of children's mental health agencies. J Emot Behav Disord. [Empirical Study; Interview; Quantitative Study]. 2011 Mar;19(1):27-40.
158. Youtz SC. Verifying the Collaboration Experience Questionnaire: Analysis of a community-campus partnership. Dissertation Abstracts International: Section B: The Sciences and Engineering. [Dissertation Empirical Study]. 1998 Jan;58(7-B):3963.
159. Carroll TL. Multidisciplinary Collaboration: A Method for Measurement. Nurs Adm Q. 1999;23(4):86-90.
160. Nuno-Solinis R, Berraondo Zabalegui I, Sauto Arce R, San Martin Rodriguez L, Toro Polanco N. Development of a questionnaire to assess interprofessional collaboration between two different care levels. Int J Integr Care. 2013 Apr;13:e015.
161. Korner M, Wirtz MA. Development and psychometric properties of a scale for measuring internal participation from a patient and health care professional perspective. BMC Health Serv Res. 2013;13:374.
162. Ushiro R. Nurse-Physician Collaboration Scale: development and psychometric testing. J Adv Nurs. 2009 Jul;65(7):1497-508.
163. Thombs BD, Adeponle AB, Kirmayer LJ, Morgan JF. A brief scale to assess hospital doctors' attitudes toward collaborative care for mental health. Can J Psychiatry. 2010 Apr;55(4):264-7.
164. Van C, Costa D, Abbott P, Mitchell B, Krass I. Community pharmacist attitudes towards collaboration with general practitioners: development and validation of a measure and a model. BMC Health Serv Res. 2012;12:320.

165. Van C, Costa D, Mitchell B, Abbott P, Krass I. Development and validation of the GP frequency of interprofessional collaboration instrument (FICI-GP) in primary care. J Interprof Care. 2012 Jul;26(4):297-304.
166. Doran JM, Safran JD, Waizmann V, Bolger K, Muran JC. The alliance negotiation scale: psychometric construction and preliminary reliability and validity analysis. Psychother Res. 2012;22(6):710-9.
167. Berendsen AJ, Groenier KH, de Jong GM, Meyboom-de Jong B, van der Veen WJ, Dekker J, et al. Assessment of patient's experiences across the interface between primary and secondary care: Consumer Quality Index Continuum of care. Patient Educ Couns. 2009 Oct;77(1):123-7.
168. Allen JG, Lewis L, Eyman JR, Coyne L. A scale to measure patient collaboration in neuropsychological assessment. J Clin Exp Neuropsychol. [Empirical Study]. 1989;11(2):66-70.
169. Vanhaecht K, De Witte K, Depreitere R, Van Zelm R, De Bleser L, Proost K, et al. Development and validation of a care process self-evaluation tool. Health Serv Manage Res. 2007 Aug;20(3):189-202.
170. Upenieks VV, Lee EA, Flanagan ME, Doebbeling BN. Healthcare Team Vitality Instrument (HTVI): developing a tool assessing healthcare team functioning. J Adv Nurs. 2010;66(1):168-76.
171. Kenaszchuk C, Reeves S, Nicholas D, Zwarenstein M. Validity and reliability of a multiple-group measurement scale for interprofessional collaboration. BMC Health Serv Res. 2010;10:83.
172. Cartwright J, Franklin D, Forman D, Freegard H. Promoting collaborative dementia care via online interprofessional education. Australas J Ageing. 2013 Oct 7.
173. Berendsen AJ, Benneker WH, Schuling J, Rijkers-Koorn N, Slaets JP, Meyboom-de Jong B. Collaboration with general practitioners: preferences of medical specialists--a qualitative study. BMC Health Serv Res. 2006;6:155.
174. Berendsen AJ, Benneker WH, Meyboom-de Jong B, Klazinga NS, Schuling J. Motives and preferences of general practitioners for new collaboration models with medical specialists: a qualitative study. BMC Health Serv Res. 2007;7:4.
175. Vandenberg RJ, Lance CE. A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. Organ Res Methods. 2000 January 1, 2000;3(1):4-70.
176. Aiken LH, Patrician PA. Measuring organizational traits of hospitals: the Revised Nursing Work Index. Nurs Res. 2000 May-Jun;49(3):146-53.
177. Raykov T. Estimation of composite reliability for congeneric measures. Appl Psych Meas. 1997;21(2):173-84.
178. Heinemann GD, Schmitt MH, Farrell MP, Brallier SA. Development of an Attitudes Toward Health Care Teams Scale. Eval Health Prof. 1999 Mar;22(1):123-42.
179. Pawson R, Tilley N. Realistic Evaluation. London: SAGE Publications Ltd; 1997.
180. Undre S, Healey A, Sevdalis N, Koutantji M, Vincent C. The Observational Teamwork Assessment for Surgery (OTAS): Development, Feasibility and Reliability. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2007 October 1, 2007;51(11):673-7.

181. Dickinson TL, McIntyre RM. A Conceptual Framework for Teamwork Measurement. In: Brannick MT, Salas E, Prince C, editors. Team Performance Assessment and Measurement. New Jersey: Lawrence Erlbaum Associates, Inc.; 1997. p. 19-44.
182. Anastasi A, Urbina S. Psychological Testing. 7th ed. New Jersey: Prentice Hall; 1997.
183. Fried GM, Feldman LS. Objective assessment of technical performance. World J Surg. 2008;32:156-60.
184. Korner M. [A model of shared decision-making in medical rehabilitation]. Rehabilitation (Stuttg). 2009 Jun;48(3):160-5.
185. Korner M, Ehrhardt H, Steger AK. Designing an interprofessional training program for shared decision making. J Interprof Care. 2013 Mar;27(2):146-54.
186. Valentine MA, Nembhard IM, Edmondson AC. Measuring Teamwork in Health Care Settings: A Review of Survey Instruments. Med Care. 2014 Apr 30.
187. West MA. Effective teamwork. Practical lessons from organization research. 3rd ed: Leicester:Blackwell.
188. Kauffeld S. Teamdiagnose. Rhein R, editor: VS Verlag für Sozialwissenschaften; 2004.
189. Kriston L, Scholl I, Holzel L, Simon D, Loh A, Harter M. The 9-item Shared Decision Making Questionnaire (SDM-Q-9). Development and psychometric properties in a primary care sample. Patient Educ Couns. 2010 Jul;80(1):94-9.
190. Scholl I, Kriston L, Dirmaier J, Buchholz A, Harter M. Development and psychometric properties of the Shared Decision Making Questionnaire--physician version (SDM-Q-Doc). Patient Educ Couns. 2012 Aug;88(2):284-90.
191. Buhrlen B, Gerdes N, Jackel WH. [Development and psychometric testing of a patient questionnaire for medical rehabilitation (IRES-3)]. Rehabilitation (Stuttg). 2005 Apr;44(2):63-74.
192. Farin E, Meixner K, Follert P, Jackel WH, Jacob A. [Job satisfaction in rehabilitation clinics--Development of the "MiZu-Reha" questionnaire and its use in quality assurance]. Rehabilitation (Stuttg). 2002 Aug;41(4):258-67.
193. GfQG - Gesellschaft für Qualität im Gesundheitswesen: Questionnaire on Patient Satisfaction (ZUF-8). Available from: http://www.gfqg.de/assessment_zuf8.pdf.
194. Odegard A, Bjorkly S. A mixed method approach to clarify the construct validity of interprofessional collaboration: an empirical research illustration. J Interprof Care. 2012 Jul;26(4):283-8.
195. Denzin NK. Triangulation 2.0. J Mix Methods Res. 2012 April 1, 2012;6(2):80-8.
196. Varela FJ, Coutinho A. Second generation immune networks. Immunol Today. 1991 May;12(5):159-66.
197. Wilson T, Holt T, Greenhalgh T. Complexity science: complexity and clinical care. BMJ. 2001 Sep 22;323(7314):685-8.
198. Engel GL. The need for a new medical model: a challenge for biomedicine. Science. 1977 Apr 8;196(4286):129-36.
199. Thurstone LL. Multiple factor analysis. Psychol Rev. 1931;38(5):406-27.

200. Thurstone LL. Multiple-factor analysis. Chicago: Wiley Subscription Services, Inc., A Wiley Company; 1947.
201. Ertel S. Exploratory factor analysis revealing complex structure. Personality and Individual Differences. 2011;50(2):196-200.
202. Eysenck HJ. Four ways five factors are not basic. Pers Individ Dif. 1992;13(6):667-73.
203. Revelle W. Factors are fictions, and other comments on individuality theory. J Pers. 1983;51(4):707-14.
204. Schonemann PH. The Psychopathology of Factor Indeterminancy. Multivariate Behav Res. 1996 1996/10/01;31(4):571-7.
205. Healey AN, Undre S, Vincent CA. Developing observational measures of performance in surgical teams. Qual Saf Health Care. 2004 Oct;13 Suppl 1:i33-40.

Appendices

Appendix 1: Critical appraisal instrument; COSMIN Checklist

The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments

COSMIN checklist with 4-point scale

Contact
CB Terwee, PhD
VU University Medical Center
Department of Epidemiology and Biostatistics
EMGO Institute for Health and Care Research
1081 BT Amsterdam
The Netherlands
Website: www.cosmin.nl, www.emgo.nl
E-mail: cb.terwee@vumc.nl

Instructions

This version of the COSMIN checklist is recommended for use in systematic reviews of measurement properties. With this version it is possible to calculate overall methodological quality scores per study on a measurement property. A methodological quality score per box is obtained by taking the lowest rating of any item in a box ('worse score counts'). For example, if for a reliability study one item in the box 'Reliability' is scored poor, the methodological quality of that reliability study is rated as poor. The Interpretability box and the Generalizability box are mainly used as data extraction forms. We recommend to use the Interpretability box to extract all information on the interpretability issues described in this box (e.g. norm scores, floor-ceiling effects, minimal important change) of the instruments under study from the included articles. Similar, we recommend to use the Generalizability box to extract data on the characteristics of the study population and sampling procedure. Therefore no scoring system was developed for these boxes.

This scoring system is described in this paper:

Terwee CB, Mokkink LB, 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 2011, July 6 [epub ahead of print].

Step 1. Evaluated measurement properties in the article

	Internal consistency	Box A
	Reliability	Box B
	Measurement error	Box C
	Content validity	Box D
	Structural validity	Box E
	Hypotheses testing	Box F
	Cross-cultural validity	Box G
	Criterion validity	Box H
	Responsiveness	Box I

Step 2. Determining if the statistical method used in the article are based on CTT or IRT

Box General requirements for studies that applied Item Response Theory (IRT) models				
	excellent	good	fair	poor
1 Was the IRT model used adequately described? e.g. One Parameter Logistic Model (OPLM), Partial Credit Model (PCM), Graded Response Model (GRM)	IRT model adequately described	IRT model not adequately described		
2 Was the computer software package used adequately described? e.g. RUMM2020, WINSTEPS, OPLM, MULTILOG, PARSCALE, BILOG, NLMIXED	Software package adequately described	Software package not adequately described		
3 Was the method of estimation used adequately described? e.g. conditional maximum likelihood (CML), marginal maximum likelihood (MML)	Method of estimation adequately described	Method of estimation not adequately described		
4 Were the assumptions for estimating parameters of the IRT model checked? e.g. unidimensionality, local independence, and item fit (e.g. differential item functioning (DIF))	assumptions of the IRT model checked	assumptions of the IRT model partly checked	assumptions of the IRT model not checked or unknown	

To obtain a total score for the methodological quality of studies that use IRT methods, the 'worse score counts' algorithm should be applied to the IRT box in combination with the box of the measurement property that was evaluated in the IRT study. For example, if IRT methods are used to study internal consistency and item 4 in the IRT box is scored fair, while the items in the internal consistency box (box A) are all scored as good or excellent, the methodological quality score for internal consistency will be fair. However, if any of the items in box A is scored poor, the methodological quality score for internal consistency will be poor.

Step 3. Determining if a study meets the standards for good methodological quality

Box A. Internal consistency				
	excellent	good	fair	poor
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?				
Design requirements				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the internal consistency analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
5 Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	Factor analysis performed in the study population	Authors refer to another study in which factor analysis was performed in a similar study population	Authors refer to another study in which factor analysis was performed, but not in a similar study population	Factor analysis NOT performed and no reference to another study
6 Was the sample size included in the unidimensionality analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 6-7* #items but <100	5* #items but <100	<5* #items
7 Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	Internal consistency statistic calculated for each subscale separately			Internal consistency statistic NOT calculated for each subscale separately
8 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
9 for Classical Test Theory (CTT), continuous scores: Was Cronbach's alpha calculated?	Cronbach's alpha calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated
10 for CTT, dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated
11 for IRT: Was a goodness of fit statistic at a global level calculated? E.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	Goodness of fit statistic at a global level calculated			Goodness of fit statistic at a global level NOT calculated

NB. Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study.

Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)				
	excellent	good	fair	poor
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Were at least two measurements available?	At least two measurements			Only one measurement
5 Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6 Was the time interval stated?	Time interval stated		Time interval NOT stated	
7 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
9 Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
11 for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12 for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13 for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14 for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box C. Measurement error: absolute measures				
	excellent	good	fair	poor
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Were at least two measurements available?	At least two measurements			Only one measurement
5 Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6 Was the time interval stated?	Time interval stated		Time interval NOT stated	
7 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
9 Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
11 for CTT: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	SEM, SDC, or LoA calculated	Possible to calculate LoA from the data presented		SEM calculated based on Cronbach's alpha, or on SD from another population

Box D. Content validity (including face validity)				
	excellent	good	fair	poor
General requirements				
1 Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	Assessed if all items refer to relevant aspects of the construct to be measured		Aspects of the construct to be measured poorly described AND this was not taken into consideration	NOT assessed if all items refer to relevant aspects of the construct to be measured

2	Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	Assessed if all items are relevant for the study population in adequate sample size (≥10)	Assessed if all items are relevant for the study population in moderate sample size (5-9)	Assessed if all items are relevant for the study population in small sample size (<5)	NOT assessed if all items are relevant for the study population OR target population not involved
3	Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	Assessed if all items are relevant for the purpose of the application	Purpose of the instrument was not described but assumed	NOT assessed if all items are relevant for the purpose of the application	
4	Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	Assessed if all items together comprehensively reflect the construct to be measured		No theoretical foundation of the construct and this was not taken into consideration	NOT assessed if all items together comprehen-sively reflect the construct to be measured
5	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Box E. Structural validity				
	excellent	good	fair	poor
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?				
Design requirements				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 5-7* #items but <100	5* #items but <100	<5* #items
5 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. rotation method not described)	Other important methodological flaws in the design or execution of the study (e.g. inappropriate rotation method)

Statistical methods					
6	for CTT: Was exploratory or confirmatory factor analysis performed?	Exploratory or confirmatory factor analysis performed and type of factor analysis appropriate in view of existing information	Exploratory factor analysis performed while confirmatory would have been more appropriate		No exploratory or confirmatory factor analysis performed
7	for IRT: Were IRT tests for determining the (uni-) dimensionality of the items performed?	IRT test for determining (uni)dimension-ality performed			IRT test for determining (uni)dimension-ality NOT performed

Box F. Hypotheses testing					
		excellent	good	fair	Poor
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
4	Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	Multiple hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
5	Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
6	Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
7	for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
8	for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)

9 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods				
10 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Box G. Cross-cultural validity				
	excellent	good	fair	poor
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	

3 Was the sample size included in the analysis adequate?	CTT: 7* #items and ≥100 IRT: ≥200 per group	CTT: 5* #items and ≥100 OR 5-7* #items but <100 IRT: ≥200 in 1 group and 100-199 in 1 group	CTT: 5* #items but <100 IRT: 100-199 per group	CTT: <5* #items IRT: (<100 in 1 or both groups
4 Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?	Both source language and target language described			Source language NOT known
5 Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages	Expertise of the translators described with respect to disease, construct, and language	Expertise of the translators with respect to disease or construct poor or not described	Expertise of the translators with respect to language not described	
6 Did the translators work independently from each other?	Translators worked independent	Assumable that the translators worked independent	Unclear whether translators worked independent	Translators worked NOT independent
7 Were items translated forward and backward?	Multiple forward and multiple backward translations	Multiple forward translations but one backward translation	One forward and one backward translation	Only a forward translation
8 Was there an adequate description of how differences between the original and translated versions were resolved?	Adequate description of how differences between translators were resolved	Poorly or NOT described how differences between translators were resolved		

9	Was the translation reviewed by a committee (e.g. original developers)?	Translation reviewed by a committee (involving other people than the translators, e.g. the original developers)	Translation NOT reviewed by (such) a committee		
10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	Translated instrument pre-tested in the target population	Translated instrument pre-tested, but unclear if this was done in the target population	Translated instrument pre-tested, but NOT in the target population	Translated instrument NOT pre-tested
11	Was the sample used in the pre-test adequately described?	Sample used in the pre-test adequately described		Sample used in the pre-test NOT (adequately) described	
12	Were the samples similar for all characteristics except language and/or cultural background?	Shown that samples were similar for all characteristics except language /culture	Stated (but not shown) that samples were similar for all characteristics except language /culture	Unclear whether samples were similar for all characteristics except language /culture	Samples were NOT similar for all characteristics except language /culture
13	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Statistical methods					
14	for CTT: Was confirmatory factor analysis performed?	Multiple-group confirmatory factor analysis performed			Multiple-group confirmatory factor analysis NOT performed
15	for IRT: Was differential item function (DIF) between language groups assessed?	DIF between language groups assessed			DIF between language groups NOT assessed

Box H. Criterion validity					
		excellent	good	fair	poor
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4	Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'

5	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
6	for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated		Correlations or AUC NOT calculated
7	for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated		Sensitivity and specificity NOT calculated

Box I. Responsiveness				
	excellent	good	fair	poor
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
5 Was the time interval stated?	Time interval adequately described			Time interval NOT described

6	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	
7	Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
Design requirements for hypotheses testing					
	For constructs for which a gold standard was not available:				
8	Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	Hypotheses formulated a priori		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
9	Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
10	Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		

11 Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)		Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
12 Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	NO information on the measurement properties of the comparator instrument(s)
13 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods				
14 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate		Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Design requirement for comparison to a gold standard For constructs for which a gold standard was available:				
15 Can the criterion for change be considered as a reasonable gold standard?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
16 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
17 for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	Correlations or Area under the ROC Curve (AUC) calculated			Correlations or AUC NOT calculated
18 for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

Interpretability

We recommend to use the Interpretability box to extract all information on the interpretability issues described in this box of the instruments under study from the included articles.

Box Interpretability	
Percentage of missing items	
Description of how missing items were handled	
Distribution of the (total) scores	
Percentage of the respondents who had the lowest possible (total) score	
Percentage of the respondents who had the highest possible (total) score	
Scores and change scores (i.e. means and SD) for relevant (sub) groups, e.g. for normative groups, subgroups of patients, or the general population	
Minimal Important Change (MIC) or Minimal Important Difference (MID)	

Generalizability

We recommend to use the Generalizability box to extract data on the characteristics of the study populations and sampling procedures of the included studies.

Box Generalisability	
Median or mean age (with standard deviation or range)	
Distribution of sex	
Important disease characteristics (e.g. severity, status, duration) and description of treatment	
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	
Countries in which the study was conducted	
Language in which the HR-PRO instrument was evaluated	
Method used to select patients (e.g. convenience, consecutive, or random)	
Percentage of missing responses (response rate)	

Appendix 2: Generalizability data extraction instrument

Generalizability

We recommend to use the Generalizability box to extract data on the characteristics of the study populations and sampling procedures of the included studies.

Total participants			
Median or mean age (with standard deviation or range)		YEARS	Standard deviation or range
Distribution of sex		% MALE	
		% FEMALE	

Composition of multidisciplinary team;

	number (or %)	%
Nurses		
Medical practitioners		
physiotherapists		
occupational therapists		
psychologists		
pharmacists		
acupuncturists		
chiropractors		
massage therapists		
podiatrists		
exercise physiologists		
naturopaths		
Chinese Medicine Practitioners		
herbalists		
homoeopaths		
social workers		
rehabilitation consultants		
employers		
workplace health and safety officers		
Other 1		
Other 2		
Other 3		
Other 4		
Patients		
Total;	0	0.00%

Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care);	
Countries in which the study was conducted;	
Language in which the instrument was evaluated;	
Method used to select patients (e.g. convenience, consecutive, or random);	
Percentage of missing responses (response rate);	

Appendix 3: Search algorithm examples

CINAHL

((MH “Psychometrics”) OR (MH “Instrument Validation”) OR (MH “Item Analysis”) OR (MH “Reliability and Validity”) OR (MH “factor analysis”) OR (TX “varimax”)) AND ((MH Attitude of Health Personnel) OR (MH “Interprofessional Relations”)) AND ((MH Collaboration) OR (TI “collaboration”) OR (AB “collaboration”) OR (MH “cooperative behavior”)) AND ((MH Questionnaires) OR (MH “Instrument Construction”) OR (AB tool) OR (AB scale) OR (AB index) OR (AB instrument) OR (AB “assess*”) OR (AB “measure*”) OR (TI tool) OR (TI scale) OR (TI index) OR (TI instrument) OR (TI “assess*”) OR (TI “measure*”))

PubMed

((Attitude of Health Personnel[mh]) OR (Interprofessional relations[mh]) OR (Cooperative behaviour[mh])) AND ((Reproducibility of results[mh]) OR (Factor analysis[mh]) OR (Factor analysis[tw]) OR (psychometrics[mh]) OR (principal component analysis[mh])) AND (collaborat*[tiab]) AND ((Questionnaires[mh]) OR (Measure*[tiab]) OR (Instrument[tiab]) OR (Tool[tiab]) OR (Scale[tiab]) OR (Index[tiab]) OR (assess*[tiab]))

Appendix 4: Excluded studies

Excluded studies and the reasons for exclusion

Study	Reason for exclusion
Aagja JP, Garg R. Measuring perceived service quality for public hospitals (PubHosQual) in the Indian context. IJPHM. 2010;4(1):60-83.	Not about collaboration measurement
Abdallah L, Fawcett J, Kane R, Dick K, Chen J. Development and psychometric testing of the EverCare Nurse Practitioner Role and Activity Scale (ENPRAS). J Am Assoc Nurse Pract. 2005;17(1):21-6.	Less than 3 participant types
Adams A, Bond S, Arber S. Development and validation of scales to measure organisational features of acute hospital wards. Int J Nurs Stud. 1995 Dec;32(6):612-27.	Not about collaboration measurement Less than 3 participant types
Allen JG, Lewis L, Eyman JR, Coyne L. A scale to measure patient collaboration in neuropsychological assessment. International Journal of Clinical Neuropsychology. [Empirical Study]. 1989;11(2):66-70.	Less than 3 participant types
Baggs J. Psychometric evaluation of collaboration and satisfaction about care decisions (CSACD) instrument. Heart & Lung. 1992;21(3):296-.	Less than 3 participant types
Baggs JG. Development of an instrument to measure collaboration and satisfaction about care decisions. J Adv Nurs. 1994 Jul;20(1):176-82.	Less than 3 participant types
Baggs JG, Schmitt MH, Mushlin AI, Mitchell PH, Eldredge DH, Oakes D, et al. Association between nurse-physician collaboration and patient outcomes in three intensive care units. Critical Care Medicine. 1999 Sep;27(9):1991-8.	Less than 3 participant types
Bagnato SJ, Neisworth JT. Collaboration and teamwork in assessment for early intervention. Child Adolesc Psychiatr Clin N Am. 1999 Apr;8(2):347-63.	Not about collaboration measurement
Barile JP, Darnell AJ, Erickson SW, Weaver SR. Multilevel measurement of dimensions of collaborative functioning in a network of collaboratives that promote child and family well-being. Am J Community Psychol. 2012 Mar;49(1-2):270-82.	Not about care delivery (about project management)
Basu S, Salisbury CL, Thorkildsen TA. Measuring collaborative consultation practices in natural environments. J Early Interv. [Empirical Study; Quantitative Study]. 2010 Mar;32(2):127-50.	Not about care delivery (about project management)
Berendsen AJ, Groenier KH, de Jong GM, Meyboom-de Jong B, van der Veen WJ, Dekker J, et al. Assessment of patient's experiences across the interface between primary and secondary care: Consumer Quality Index Continuum of care. Patient Educ Couns. 2009 Oct;77(1):123-7.	Less than 3 participant types
Berg CA, Schindler I, Maharajh S. Adolescents' and Mothers' Perceptions of the Cognitive and Relational Functions of Collaboration and Adjustment in Dealing With Type 1 Diabetes. J Fam Psychol. 2008	Not about care delvery. Less than 3 participant types

;22(6):865-74.	
Boyer L, Belzeaux R, Maurel O, Baumstarck-Barrau K, Samuelian J-C. A social network analysis of healthcare professional relationships in a French hospital. Int J Health Care Qual Assur. 2010;23(5):460-9.	Not about the measurement of collaboration but evaluation of social networking, which has some relevance to collaboration.
Brock KA, Doucette WR. Collaborative working relationships between pharmacists and physicians: an exploratory study. JAPhA. 2004 2004;44(3):358-65.	Less than 3 participant types
Bronstein LR. Index of interdisciplinary collaboration. Soc Work Res. [Empirical Study]. 2002 Jun;26(2):113-26.	Less than 3 participant types
Capafons A, Espejo B, Mendoza M. Confirmatory factor analysis of the Valencia Scale on Attitudes and Beliefs toward Hypnosis-Therapist version. Int J Clin Exp Hypn. [Empirical Study; Quantitative Study]. 2008 Jul-Sep;56(3):281-94.	Less than 3 participant types
Chow AY. Investigating and measuring motivation in collaborative inquiry-based project settings. Dissertation Abstracts International Section A: Humanities and Social Sciences. [Dissertation Empirical Study; Quantitative Study]. 2009;69(11-A):4241.	About project collaboration not care delivery
Carroll TL. Multidisciplinary Collaboration: A Method for Measurement. Nurs Adm Q. 1999;23(4):86-90.	Abstract only
Curran V, Hollett A, Casimiro LM, McCarthy P, Banfield V, Hall P, et al. Development and validation of the interprofessional collaborator assessment rubric (ICAR). J Interprof Care. 2011;25(5):339-44.	About interprofessional education
D'Amour D, Goulet L, Labadie J-F, San Martin-Rodriguez L, Pineault R. A model and typology of collaboration between professionals in healthcare organizations. BMC Health Serv Res. 2008 Sep 21;8.	Qualitative study, no quantitative validity data
Doran JM, Safran JD, Waizmann V, Bolger K, Muran JC. The alliance negotiation scale: psychometric construction and preliminary reliability and validity analysis. Psychother Res. 2012;22(6):710-9.	Patient self reporting (one participant type)
Dougherty MB, Larson E. A review of instruments measuring nurse-physician collaboration. J Nurs Adm. 2005;35(5):244-53.	Review
Dougherty MB, Larson EL. The nurse-nurse collaboration scale. J Nurs Adm. 2010 Jan;40(1):17-25.	Less than 3 participant types
Duckers ML, Wagner C, Groenewegen PP. Developing and testing an instrument to measure the presence of conditions for successful implementation of quality improvement collaboratives. BMC Health Serv Res. 2008;8:172.	Not regarding care delivery
Dunleavy KN, Martin MM. A Convergent Validity Study of the Decision-Making Collaboration Scale. North American Journal of Psychology. [Empirical Study; Quantitative Study]. 2006;8(2):339-44.	Not healthcare related
Elg M, Stenberg J, Kammerlind P, Tullberg S, Olsson J. Swedish healthcare management practices and quality improvement work: development trends. Int J Health Care Qual Assur. 2011;24(2):101-23.	About healthcare management.
El-Zubeir M, Rizk DE, Al-Khalil RK. Are senior UAE medical and nursing	About IPE

students ready for interprofessional learning? Validating the RIPL scale in a Middle Eastern context. J Interprof Care. 2006;20(6):619-32.	
Ferrie S, Allman-Farinelli M. Development of a tool to measure dietitians' involvement in the intensive care setting. Nutr Clin Pract. 2011 Jun;26(3):330-8.	Dietician's only; ie less than 3 participant types
Foy R, Hempel S, Rubenstein L, Suttorp M, Seelig M, Shanman R, et al. Meta-analysis: effect of interactive communication between collaborating primary care physicians and specialists. Ann Intern Med. 2010 Feb 16;152(4):247-58.	Systematic review
Gross CJ. Development of an instrument to measure collaborative competencies in interprofessional health care education: North Dakota State University; 2012. Thesis	About IPE
Guevara JP, Greenbaum PE, Shera D, Shea JA, Bauer L, Schwarz DF. Development and psychometric assessment of the collaborative care for attention-deficit disorders scale. Ambul Pediatr. 2008 Jan-Feb;8(1):18-24.	Pediatricians only; ie less than 3 participant types
Hall DJ, Skipper JB, Hazen BT, Hanna JB. Inter-organizational IT use, cooperative attitude, and inter-organizational collaboration as antecedents to contingency planning effectiveness. International Journal of Logistics Management, The. 2012;23(1):50-76.	About management, not healthcare related
Hall P, Marshall D, Weaver L, Boyle A, Taniguchi A. A Method to Enhance Student Teams in Palliative Care: Piloting the McMaster-Ottawa Team Observed Structured Clinical Encounter. J Palliat Med. 2011;14(6):744-50.	About IPE
Healey AN, Undre S, Vincent CA. Developing observational measures of performance in surgical teams. Qual Saf Health Care. 2004 Oct;13 Suppl 1:i33-40.	Development of observational assessment of surgical team activity, not a validation study
Heinemann GD, Schmitt MH, Farrell MP, Brallier SA. Development of an Attitudes Toward Health Care Teams Scale. Eval Health Prof. 1999 Mar;22(1):123-42.	A measurement of attitudes towards the team, not necessarily collaboration
Henneman EA, Kleppel R, Hinchey KT. Development of a checklist for documenting team and collaborative behaviors during multidisciplinary bedside rounds. J Nurs Adm. 2013 May;43(5):280-5.	Not clear as to the participant mix ? nurses and physicians only
Hojat M, Herman MW. Developing an instrument to measure attitudes toward nurses: preliminary psychometric findings. Psychol Rep. 1985 Apr;56(2):571-9.	Students used as sample.
Hojat M, Fields SK, Veloski J, Griffiths M, Cohen MJ, Plumb JD. Psychometric properties of an attitude scale measuring physician-nurse collaboration. Eval Health Prof. [Empirical Study]. 1999 Jun;22(2):208-20.	Less than 3 participant types; nurses only
Hojat M, Nasca TJ, Cohen MJM, Fields SK, Rattner SL, Griffiths M, et al. Attitudes toward physician-nurse collaboration: A cross-cultural study of male and female physicians and nurses in the United States and Mexico. Nurs Res. 2001 Mar-Apr;50(2):123-8.	Less than 3 participant types; physicians and nurses

Hollins NL, Townsend SC. Opportunity model of collaboration: a model for assessment instrument development. J Allied Health. 2003;32(4):221-6.	Model development, not validation study etc
Hyrkäs K, Appelqvist-Schmidlechner K, Paunonen-Ilmonen M. Translating and validating the Finnish version of the Manchester Clinical Supervision Scale. Scand J Caring Sci. 2003;17(4):358-64.	About clinical supervison, not care delivery; translation of instrument checking cross-cultural validity.
Kenaszchuk C, Conn LG, Dainty K, McCarthy C, Reeves S, Zwarenstein M. Consensus on interprofessional collaboration in hospitals: statistical agreement of ratings from ethnographic fieldwork and measurement scales. J Eval Clin Pract. 2012 Feb;18(1):93-9.	One participant type only; nurses
Killaspy H, White S, Taylor TL, King M. Psychometric properties of the Mental Health Recovery Star. Br J Psychiatry. [Empirical Study; Quantitative Study]. 2012 Jul;201(1):65-70.	Tool aims to assess a person's recovery from mental ill health, not collaboration
Konrad TR, Fletcher GS, Carey TS. Interprofessional collaboration and job satisfaction of chiropractic physicians. J Manipulative Physiol Ther. 2004 May;27(4):245-52.	One participant type only; chiropractors
Lauffs M, Ponzer S, Saboonchi F, Lonka K, Hylin U, Mattiasson AC. Cross-cultural adaptation of the Swedish version of Readiness for Interprofessional Learning Scale (RIPLS). Med Educ. 2008 Apr;42(4):405-11.	About IPE, trans-cultural translation
Le Q, Spencer J, Whelan J. Development of a tool to evaluate health science students' experiences of an interprofessional education (IPE) programme. Ann Acad Med Singapore. 2008 Dec;37(12):1027-33.	About IPE
Legare F, Moher D, Elwyn G, LeBlanc A, Gravel K. Instruments to assess the perception of physicians in the decision-making process of specific clinical encounters: a systematic review. BMC Med Inform Decis Mak. 2007;7:30.	Systematic review; not validation study
Lim KH. Collaboration between disciplinary teams caring for elders in Korean community settings: University of Arizona; 2008.	Less than 3 participant types; social workers and nurses
Lindhardt T, Nyberg P, Hallberg IR. Collaboration between relatives of elderly patients and nurses and its relation to satisfaction with the hospital care trajectory. Scand J Caring Sci. 2008;22(4):507-19.	1 participant type only; relatives
Lindhardt T, Nyberg P, Hallberg IR. Relatives' view on collaboration with nurses in acute wards: development and testing of a new measure. Int J Nurs Stud. 2008 Sep;45(9):1329-43.	1 participant type only; relatives
Lindhardt T, Nyberg P, Hallberg IR. Relatives' view on collaboration with nurses in acute wards: development and testing of a new measure. Int J Nurs Stud. 2008 Sep;45(9):1329-43.	1 participant types only; relatives
Liu Y, Doucette WR, Farris KB. Examining the development of pharmacist-physician collaboration over 3 months. Res Social Adm Pharm. 2010 Dec;6(4):324-33.	Less than 3 participant types
Lockyer JM, Violato C, Fidler H. A multi source feedback program for anesthesiologists. Can J Anaesth. 2006;53(1):33-9.	More about assessment of anesthesiologists performance than about collaboration

Martin-Rodríguez LS, D'Amour D, Leduc N. Validation of an intensity of interprofessional collaboration questionnaire translated into Spanish [Spanish]. Enferm Clin. 2007 2007 Jan-Feb;17(1):24-31.	Less than 3 participant types; nurses only
Masse LC, Moser RP, Stokols D, Taylor BK, Marcus SE, Morgan GD, et al. Measuring collaboration and transdisciplinary integration in team science. Am J Prev Med. 2008 Aug;35(2):S151-S60.	About team science, not care delivery
Maylone MM, Ranieri L, Quinn Griffin MT, McNulty R, Fitzpatrick JJ. Collaboration and autonomy: perceptions among nurse practitioners. J Am Acad Nurse Pract. 2011 Jan;23(1):51-7.	Less than 3 participant types; nurses only
Melin A, Granath J-Å. Patient focused healthcare: an important concept for provision and management of space and services to the healthcare sector. Facilities. 2004;22(11):284-9.	About model development, not validation
Nansel TR, Rovner AJ, Haynie D, Iannotti RJ, Simons-Morton B, Wysocki T, et al. Development and validation of the collaborative parent involvement scale for youths with type 1 diabetes. J Pediatr Psychol. 2009 Jan-Feb;34(1):30-40.	Less than 3 participant types; youths only
Nierenberg AA, Ostacher MJ, Borrelli DJ, Iosifescu DV, Perlis RH, Desrosiers A, et al. The integration of measurement and management for the treatment of bipolar disorder: A STEP-BD model of collaborative care in psychiatry. J Clin Psychiatry. 2006 //;67(SUPPL. 11):3-7.	About model development, not validation study
Odegard A, Bjorkly S. A mixed method approach to clarify the construct validity of interprofessional collaboration: an empirical research illustration. J Interprof Care. 2012 Jul;26(4):283-8.	Report of model for triangulation mixed methods study for IPC, not a validation study
oi J, Bakken S, Larson E, Du Y, Stone PW. Perceived nursing work environment of critical care nurses. Nurs Res. 2004 Nov-Dec;53(6):370-8.	Less than 3 types of participants; ICU nurses only
Padma P, Rajendran C, Sai LP. A conceptual framework of service quality in healthcare: Perspectives of Indian patients and their attendants. Benchmarking: An International Journal. 2009;16(2):157-91.	About service quality, less than 3 participant types
Paris M, Bedregal LE, Anez LM, Shahar G, Davidson L. Psychometric Properties of the Spanish Version of the Therapeutic Collaboration Scale (TCS). Hisp J Behav Sci. [Empirical Study; Quantitative Study]. 2004 Aug;26(3):390-402.	Less than 3 participant types; patients only
Parker K, Jacobson A, McGuire M, Zorzi R, Oandasan I. How to build high-quality interprofessional collaboration and education in your hospital: the IP-COMPASS tool. Qual Manag Health Care. 2012 2012;21(3):160-8.	About IPE.
Parsell G, Bligh J. The development of a questionnaire to assess the readiness of health care students for interprofessional learning (RIPLS). Med Educ. 1999;33(2):95-100.	About IPE
Pehl LKH. Development of an instrument to measure perceptions of collaboration between nursing deans and nursing service administrators: UNIVERSITY OF TEXAS AT AUSTIN; 1988.	Less than 3 participant types; nursing deans and administrators. Not care delivery related.
Petri L. Concept analysis of interdisciplinary collaboration. Nurs Forum.	Concept analysis not instrument

2010 2010;45(2):73-82.	development and validation
Polivka BJ, Dresbach SH, Heimlich JE, Elliott M. Interagency relationships among rural early intervention collaboratives. Public Health Nurs. 2001 Sep-Oct;18(5):340-9.	Interagency collaboration measurement, not care related
Reeb RN, Folger SF, Oneal BJ. Behavioral Summarized Evaluation: An assessment tool to enhance multidisciplinary and parent-professional collaborations in assessing symptoms of autism. Child Health Care. [Empirical Study; Quantitative Study]. 2009 Oct;38(4):301-20.	Review of validation studies. About assessing autism not necessarily about assessing collaboration
Remneland-Wikhamn B, Wikhamn W. Open innovation climate measure: The introduction of a validated scale. Creativity and Innovation Management. [Empirical Study; Interview; Quantitative Study]. 2011 Dec;20(4):284-95.	Not healthcare
San Martin-Rodriguez L, D'Amour D, Leduc N. Validation of an intensity of interprofessional collaboration questionnaire translated into Spanish. Enferm Clin. 2007;17(1):24-31.	Not in English
Scholl I, Kriston L, Dirmaier J, Buchholz A, Härter M. Development and psychometric properties of the Shared Decision Making Questionnaire - physician version (SDM-Q-Doc). Patient Educ Couns. 2012;88(2):284-90.	Less than 3 participant types; physicians and patients
Sexton JB, Holzmueller CG, Pronovost PJ, Thomas EJ, McFerran S, Nunes J, et al. Variation in caregiver perceptions of teamwork climate in labor and delivery units. J Perinatol. 2006;26(8):463-70.	About attitudes towards safety
Sexton JB, Makary MA, Tersigni AR, Pryor D, Hendrich A, Thomas EJ, et al. Teamwork in the operating room: Frontline perspectives among hospitals and operating room personnel. Anesthesiology. 2006;105(5):877-84.	About attitudes towards safety
Shields CG, Franks P, Fiscella K, Meldrum S, Epstein RM. Rochester Participatory Decision-Making Scale (RPAD): Reliability and Validity. Ann Fam Med. [Empirical Study; Quantitative Study]. 2005 Sep-Oct;3(5):436-42.	Less than 3 participant types
Shortell SM, Rousseau DM, Gillies RR, Devers KJ, Simons TL. Organizational assessment in intensive care units (ICUs): construct development, reliability, and validity of the ICU nurse-physician questionnaire. Med Care. 1991 Aug;29(8):709-26.	Less than 3 participant types; physicians and nurses
Sicotte C, D'Amour D, Moreault MP. Interdisciplinary collaboration within Quebec community health care centres. Soc Sci Med. 2002 //;55(6):991-1003.	Not validation study
Siedlecki SL, Hixson ED. Development and psychometric exploration of the professional practice environment assessment scale. J Nurs Scholarsh. 2011 Dec;43(4):421-5.	Less than 3 participant types; physicians and nurses
Steinheider B, Bayerl PS, Menold N, Bromme R. Development and validation of a scale to assess knowledge integration problems in interdisciplinary project teams (WIP). Zeitschrift fur Arbeits- und Organisationspsychologie. [Empirical Study; Qualitative Study; Quantitative Study]. 2009;53(3):121-30.	About interdisciplinary project teams and not about care delivery

Strating MM, Nieboer AP. Norms for creativity and implementation in healthcare teams: testing the group innovation inventory. Int J Qual Health Care. 2010 Aug;22(4):275-82.	About measuring group innovation; combines aspects of collaboration with other parameters. Excluded based on the perception that this tool measures more than just collaboration.
Tamura Y, Seki K, Usami M, Taku S, Bontje P, Ando H, et al. Cultural adaptation and validating a Japanese version of the readiness for interprofessional learning scale (RIPLS). J Interprof Care. 2012 Jan;26(1):56-63.	About IPE
Tan K, Chan M, Lim W, Baharom Adzahar F, Lim I. Transactive memory system as a measure of interprofessional collaboration in a multidisciplinary geriatric team. J Am Geriatr Soc. 2013;61:S57.	A measure of knowledge and expertise within MDRs. Translates as a measure of IPC, but it is unclear if this is measuring collaboration or just a predictor of collaboration.
Teixeira de Melo AMI. Validity and Reliability of Three Rating Scales to Assess Practitioners' Skills to Conduct Collaborative, Strength-Based, Systemic Work in Family-Based Services. Am J Fam Ther. [Article]. 2012 10//Oct-Dec2012;40(5):420-33.	Rating scale that assesses practitioner's skills and knowledge, not a specifiic measure of collaboration.
Thannhauser J, Russell-Mayhew S, Scott C. Measures of interprofessional education and collaboration. J Interprof Care. 2010;24(4):336-49.	Review
Thombs BD, Adeponle AB, Kirmayer LJ, Morgan JF. A brief scale to assess hospital doctors' attitudes toward collaborative care for mental health. Can J Psychiatry. 2010 Apr;55(4):264-7.	Less than 3 participant types
Undre S, Healey A, Sevdalis N, Koutantji M, Vincent C. The Observational Teamwork Assessment for Surgery (OTAS): Development, Feasibility and Reliability. Proc Hum Fact Ergon Soc Annu Meet. 2007 October 1, 2007;51(11):673-7.	Development and feasibility study only. Not a validation study.
Ushiro R. Nurse-Physician Collaboration Scale: development and psychometric testing. J Adv Nurs. 2009 Jul;65(7):1497-508.	Less than 3 participant types
Van C, Costa D, Abbott P, Mitchell B, Krass I. Community pharmacist attitudes towards collaboration with general practitioners: development and validation of a measure and a model. BMC Health Serv Res. 2012;12:320.	Less than 3 participant types
Van C, Costa D, Mitchell B, Abbott P, Krass I. Development and validation of the GP frequency of interprofessional collaboration instrument (FICI-GP) in primary care. J Interprof Care. 2012 Jul;26(4):297-304.	Less than 3 participant types
Van C, Costa D, Mitchell B, Abbott P, Krass I. Development and initial validation of the Pharmacist Frequency of Interprofessional Collaboration Instrument (FICI-P) in primary care. Res Social Adm Pharm. 2012 //;8(5):397-407.	Less than 3 participant types
Weiss SJ, Davis HP. Validity and reliability of the Collaborative Practice Scales. Nurs Res. 1985 Sep-Oct;34(5):299-305.	Less than 3 participant types

Wilhelmsson M, Ponzer S, Dahlgren LO, Timpka T, Faresjo T. Are female students in general and nursing students more ready for teamwork and interprofessional collaboration in healthcare? BMC Med Educ. 2011;11:15.	About IPE
Williams B, Brown T, Boyle M. Construct validation of the readiness for interprofessional learning scale: a Rasch and factor analysis. J Interprof Care. 2012 Jul;26(4):326-32.	About IPE
Yildirim A, Akinci F, Ates M, Ross T, Issever H, Isci E, et al. Turkish version of the Jefferson Scale of Attitudes Toward Physician-Nurse Collaboration: a preliminary study. Contemp Nurse. 2006 Oct;23(1):38-45.	Less than 3 participant types
Youtz SC. Verifying the Collaboration Experience Questionnaire: Analysis of a community-campus partnership. Dissertation Abstracts International: Section B: The Sciences and Engineering. [Dissertation Empirical Study]. 1998 Jan;58(7-B):3963.	About community health projects not healthcare delivery
Zillich AJ, Doucette WR, Carter BL, Kreiter CD. Development and initial validation of an instrument to measure physician-pharmacist collaboration from the physician perspective. Value Health. 2005 Jan-Feb;8(1):59-66.	Less than 3 participant types
Zillich AJ, Milchak JL, Carter BL, Doucette WR. Utility of a questionnaire to measure physician-pharmacist collaborative relationships. J Am Pharm Assoc. 2006 Jul-Aug;46(4):453-8.	Less than 3 participant types

Appendix 5: The characteristics of the included studies table

Authors	Instrument	Language	Sample characteristics	Measurement properties (methodological quality), interpretability	Summary findings
Oliver, Wittenberg-Lyles et al. (2006)	**MIIC**	English	Nurses (n=41) Social workers (n=18) Chaplains (n=9) Administration (n=6) Other clinical (n=13) Unknown (n=8) **Setting** Hospice 1; free standing average sized urban unit California, USA. Hospice 2; free standing, average sized rural unit New York, USA. Hospice 3; hospital based, average sized rural unit Missouri, USA. Hospice 4; hospital based small rural unit Missouri, USA. Hospice 5; free standing large urban unit Nevada, USA.	Interpretability; Means and SD for the four subscales and total scale reported for each of five different hospices. ANOVA was performed to assess differences between the 5 programs.	"...one-way ANOVA... did not find any statistically significant differences between disciplines..." " ANOVA ... of the five hospice programs did ... reveal significant differences on the mean total instrument score and on three of four subscales." The MIIC may be useful to gauge the effect of implemented programs aimed at improving collaborative care.

Oliver, Wittenberg-Lyles et al. (2007)	**MIIC**	English	Nurses (n=41) Social workers (n=18) Chaplains (n=9) Administration (n=6) Other clinical (n=13) Unknown (n=8) **<u>Setting</u>** Hospice 1; free standing average sized urban unit California, USA. Hospice 2; free standing, average sized rural unit New York, USA. Hospice 3; hospital based, average sized rural unit Missouri, USA. Hospice 4; hospital based small rural unit Missouri, USA. Hospice 5; free standing large urban unit Nevada, USA.	Internal consistency (fair) Reference to another study only for factor analysis but not a similar study population. Content validity (good) Items relevant and reflect the construct. Criterion validity (poor) Correlations not calculated. Interpretability; Means and SD for the four subscales and total scale reported.	The MIIC is comparable psychometrically to the original instrument (IIC), however further evidence is needed.

Wittenberg-Lyles and Parker Oliver (2007)	**MIIC**	English	Nurses (n=8) Social workers (n=3) Chaplains (n=2) Administration (n=1) Other clinical (n=2) Unknown (n=2) **Setting** Large, urban, free-standing hospice agency USA.	Interpretability; Range, means and SD for the four subscales reported.	"…findings suggest that interdisciplinary collaboration also occurs outside of hospice, namely with primary care doctors and nursing home staff." A mixed quantitative-qualitative design adding to the interpretability of the MIIC.
Wittenberg-Lyles, Parker Oliver et al. (2010)	**MIIC**	English	Completing the MIIC Nurses (n=12) Social workers (n=2) Chaplains (n=2) Other (n=2) Unknown (n=2) **Setting** One hospice Midwestern USA.	Interpretability; Means and SD for the four subscales and total scale for both teams and all teams reported.	"…team's reflection on process was most likely to occur in team meetings, … least likely to occur when caregivers were present… team members had a high perception of interdependence and flexibility of roles… less likely to be enacted in team meetings..." "Caregiver participation in team meetings had a positive impact on collaborative communication…" Study demonstrates perception of collaboration measured by the MIIC diverges from the actual collaborative acts.

Berendsen, Benneker et al. (2010)	DOC	English	<u>Physicians (n=136)</u> Psychiatrist (n=30) Internist (n=29) Paediatrician (n=28) Cardiologist (n=12) Neurologist (n=12) Rehabilitation doctor (n=9) Pulmonologist (n=6) Dermatologist (n=5) Clinical geriatrician (n=3) Allergologist (n=1) Rheumatologist (n=1) <u>Surgeons (n=70)</u> Ophthalmic surgeon (n=16) Gynaecologist (n=14) General surgeon (n=11) Urologist (n=8) Orthopaedic surgeon (n=5)	Content validity (excellent) Internal consistency (fair) Structural validity (fair) Hypothesis testing (fair) Methodological quality suffers from a lack of reporting re missing items.	For assessing collaboration between doctors and specialists the DOC questionnaire is a useful tool. It is unlikely to be applicable to measuring collaboration in situations where the participant mix is more complex beyond GPs and specialists.

			Orofacial surgeon (n=4) Plastic surgeon (n=4) ENT doctor (n=6) Thoracic surgeon (n=2) <u>Support specialists (n=25)</u> Radiologist (n=11) Radio therapist (n=7) Microbiologist (n=4) Pathologist (n=3) **<u>Setting</u>** GPs & specialists in the Netherlands.		
Kenaszchuk, Reeves et al. (2010)	**MGMS**	English	Nurses (n=479) Physicians (n=127) Allied professionals (n=217) **<u>Setting</u>** 15 community and teaching hospitals in Canada.	Internal consistency (good) Structural validity (good) Hypothesis testing (good) Criterion validity (good) Methodological quality suffers from lack of reporting re missing items.	The MGMS demonstrates potential for the rating of collaboration by multiple groups rating each other.

King, Shaw et al. (2010)	**ISVS**	English	Nursing (RN) (n=26) Nursing (RPN) (n=2) Medicine (n=7) Physical therapy (n=11) Occupational therapy (n=38) Psychology (n=3) Social work (n=3) Dietetics (n=1) Clinical kinesiology(n=1) Pre-professional program (n=8) Speech language pathology(n=1) Other (n=9) Missing (n=14) **Setting** Students in CIPHER-MH project Canada	Internal consistency (fair) Content validity (excellent) Structural validity (fair) Methodological quality suffers from lack of reporting re missing items.	Application for the assessment of socio-cultural aspects of collaborative practice. Limited generalizability from this study due sample predominantly of occupational therapy and nursing students.
Nuno-Solinis, Berraondo Zabalegui et al. (2013)	**IPC-DLC**	Spanish	n=187 Primary care nurses (43%) GPs paediatricians (31%)	Internal consistency (fair) Content validity (excellent)	Useful in measuring collaboration in a sample of doctors and nurses from different levels of healthcare, but may have expanded utility with other participant types.

		Hospital specialists (18.5%) Hospital nurses (6%) **<u>Setting</u>** Clinical professionals working in the Basque Health Service in Spain.	Structural validity (fair) Methodological quality suffers from lack of reporting re missing items.	
Mellin, Bronstein et al. (2010)	**IITC-ESMH**	English	Nurses (n=68) Psychologists (n=64) Social workers (n=145) Counselling (n=78) Other (n=81) **<u>Setting</u>** Interprofessional teams in schools in the USA.	Internal consistency (excellent) Content validity (excellent) Structural validity (excellent)	This study is a significant contribution to the validity of the IIC/Bronstein model and is useful in measuring team function in Expanded School Mental Health.

Orchard, King et al. (2012)	**AITCS**	English	n=125 Registered Nurses (58.5%) Medical practitioners (2.5%) Physiotherapists (8.5%) Occupational therapists (5.1%) Pharmacists (4.2%) Social workers (5.9%) Dieticians (2.5%) Practice nurse (2.5%) Other (10.3%) **<u>Setting</u>** 7 healthcare teams from orthopedic, general surgery, acute mental health, palliative care in Canada.	Content validity (excellent) Internal consistency (poor) Sample size inadequate for unidimensionality analysis [#] Structural validity (poor) Sample size inadequate for factor analysis[#] # sample is less than five times the number of items (COSMIN standard)	Addresses collaborative relationships in the healthcare setting and has a potential for use with collaborative teams with diverse participant types.

Schroder, Medves et al. (2011)	**CPAT**	English	Nurses (n=32) Practice nurse (n=16) Medical practitioners(n=6) Physiotherapists(n=6) Occupational therapists (n=7) Pharmacists(n=3) Social workers(n=6) Dietician (n=4) Spiritual care (n=3) Other (n=28) **Setting** A palliative care team, a geriatric assessment team and two family practice teams in Canada.	Content validity (n=42) (excellent) Internal consistency (poor) Sample size inadequate for unidimensionality analysis [#] [#] sample is less than five times the number of items (COSMIN standard)	Utility with a complex mix of participant types to assess collaboration as an assessment for team function.
Upenieks, Lee et al. (2010)	**HTVI**	English	**Phase 1 Interviews** Nurses (n=15) Physiotherapists (n=1) Respiratory therapists (n=2) **Phase 1 Criterion validity** n= 439 Registered Nurses (54%)	Content validity (fair) Criterion validity (fair) Structural validity (fair) Methodological quality suffers from lack of reporting re missing items. Minor flaws in study design; minimal	Useful in measuring team collaboration in hospital units.

	Assistive personnel (21%) Licensed Vocational Nurse (10%) Unit secretaries (6%) Physicians (2%) Other training backgrounds (7%) **Phase 2 Structural validity** n= 464 Registered Nurses (52%) Assistive personnel (23%) Licensed Vocational Nurse (12%) Unit secretaries (6%) Physicians (1%) Other training backgrounds (6%) **Setting** Various medical-surgical units in the USA.	participation by physicians; volunteers only.

Vanhaecht, De Witte et al. (2007)	CPSET	**English**	Medical doctors (n=117) Nurses (n=151) Allied health (n=111) Pathway coordinators (n=132) **<u>Content validity</u>** n=50 Nurses (n=11) Medical practitioners (n=5) Allied health professionals (n=7) Senior hospital managers (n=7) Support dept e.g. lab, radiology etc (n=5) Doctors in hospitals (n=7) Patients (n=8**)** **<u>Setting</u>** Belgium - Dutch Clinical Pathway Network; Belgium and Netherlands**.**	**Content validity (excellent)** **Internal consistency (excellent)** **Reliability (good)** **Structural validity (excellent)** **Criterion validity (good)**	**For evaluating care processes, with a significant component being team collaboration.**

Face validity

n=83

Setting

Six multidisciplinary teams from Belgium and the Netherlands.

Reliability, construct and criterion validity

n=511

Medical doctors (n=117)

Nurses (n=151)

Allied health professionals (n=111)

Pathway coordinators (n=132)

Setting

54 different organizations (acute hospitals and rehabilitation centres) participated in multicentre study with 142 care processes from 17 different clinical areas.

Seys, Deneckere et al. (2013)	**CPSET**	English	Nurses (n=1719) Medical practitioners (n=543) Paramedics (n=524) Coordinators (n=134) Others (n=219) **Setting** Acute, psychiatric, specialized hospitals and primary care in Belgium and the Netherlands.	Internal consistency (good) Structural validity(excellent)	Creating a table from CPSET scores is proposed to be helpful for healthcare managers to rank teams in their facility.
Korner and Wirtz (2013)	**IPS**	German	Patients (n=536) Nurses (n=48) Medical practitioners (n=49) Physiotherapists (n=50) Psychosocial therapists (n=67) Others (n=37) More than one professional group (n=12) Missing (n=9) **Setting** Somatic and psychosomatic rehabilitation clinics in Germany.	Internal consistency (good) Content validity (poor) Structural validity (good) Criterion validity (good)	A short instrument for measuring internal participation with interprofessional teams and patients.

Undre, Sevadalis et al. OTAS (2007)	English	**Participants (Operating team)** Surgeons Nurses Anaesthesiologists **Participants (Observing team)** Urological surgeon Psychologist **Setting** One teaching hospital and one specialist treatment centre in the U.K. Teams were consistent across sites. 50 urology procedures, mix types of operations e.g. cystoscopy, ureteroscopy, ureterorenoscopy, transurethral resection of the prostate, orchidectomy, vasectomy, circumcision.	Interpretability; Means and SD for all phases of surgery, all behaviors and for all three OR participant types reported. Reliability – interrater (fair) Methodological quality suffers from a lack of reporting re missing items.	Anaesthetists were low on communication, nurses low on both communication and leadership, surgeons were low on communication and their scores deteriorated towards the end of the operations, affecting all of a surgeon's behaviors except coordination. This study contributes evidence for the reliability of the OTAS instrument applicable to the assessment of OR team behaviors in urological surgery.
Sevdalis, Lyons et al. OTAS (2009)	English	**Participants (Operating team)** Surgeons Nurses Anaesthesiologists **Participants (Observing team)** Expert 1 (psychologist)	Hypothesis test (fair)	Empirical support for 2 hypothesis obtained; $H_{1a;}$ Stronger correlations between Expert 1 and Expert 2 than between Expert 1 and Novice scores. $H_{1b;}$ Significant differences between Expert 1 and Novice scores; fewer or no differences between Expert 1 and Expert

			Expert 2 (psychologist) Novice (human factors - observation) **Setting** Two London teaching hospitals. 12 elective urology procedures, mix types of operations e.g. cystoscopy, ureteroscopy, ureterorenoscopy, transurethral resection of the prostate, orchidectomy, vasectomy, circumcision.		2 scores. The result of this study provides limited evidence of construct validity of the OTAS instrument.
Russ, Hull et al. (2012)	**OTAS**	English	**Participants (Operating team)** Surgeons Nurses Anaesthesiologists **Participants (Observing team)** OTAS expert (psychologist) 4 trainees (2 psychologists, 2 surgeons) **Setting** One London teaching hospitals. 14 general surgical procedures e.g. laparoscopic/open hernia repair, laparoscopic /open cholecystectomies, staging	Reliability (fair) Interpretability	Interrater reliability used as a measure of improved rater performance. The OTAS was used to measure rater performance over three phases of training compared to the performance of an expert rater.

			laparoscopies, hemicholectomies, laparoscopic appendicectomies, laparoscopic fundoplications.		
O'Leary, Boudreau et al. (2012)	**OTAS**	English	**Participant (Structured Interdisciplinary Rounds) Subteams** Physicians Nurses Social workers/case managers Pharmacists Coleaders **Participants (Observing team)** Medical librarian (observational researcher) Medical Researcher **Setting** Northwestern Memorial Hospital, Chicago USA. 7-8 independent observations on each unit (n=44 total observations) Joint observations for interrater reliability (n=20)	Reliability (poor) Methodological quality suffers from low sample numbers. Interpretability	Interrater reliability across units was excellent , across collaborative domains was good and across subteams was good to excellent, except physicians which was poor. The OTAS was used to measure collaborative teamwork in hospital teams performing structured rounds. This study showed significant differences in teamwork scores across units and collaborative domains and borderline differences across subteams

Passauer-Baierl, Hull et al. (2014)	**OTAS-D**	English	**<u>Stage 1 - systematic translation of the OTAS from English to German</u>** Psychologist trained in OTAS; forward translation of OTAS into German. General surgeon; backwards translation of OTAS to English. Expert OTAS developer; review of back-translation Expert psychologist observer; checked revised German version of the OTAS for clarity and comprehension. **<u>Stage 2 – semi-structured interviews</u>** 9 OR experts from 3 different hospitals; 3 surgeons 3 nurses (2 OR and 1anesthesia) 3 anaesthesiologists **<u>Stage 3 – prospective observational study in German OR using OTAS-D</u>** **<u>Participants (Operating room)</u>** Surgeons Nurses Anaesthesiologists **<u>Participants (Observing team)</u>** 2 blinded independent expert OTAS raters	Cross cultural validity (fair) Content validity (good) Reliability (fair)	OTAS systematically translated from English to German Content thoroughly validated via semi-structured interviews and thematic analysis. Interrater agreement and consistency between expert raters adequately evidenced. The OTAS-D is a valid and reliable instrument for assessing teamwork behaviors in the German OR setting.

	Setting 11 randomly selected surgical procedures 10 general surgical procedures; 4 cholecystectomies 3 hemicolectomies 1 appendectomy 1 parathyroidectomy 1 fundopliccation. One vascular procedure. Location of procedures is not clear.