Evaluation and Program Planning 60 (2017) 265–276

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

The use of concept mapping in measurement development and evaluation: Application and future directions

Scott R. Rosas[a,*], John W. Ridings[b]

[a] Concept Systems, Inc., 136 East State Street, Ithaca, NY 14850, United States
[b] The Institute for Clinical Social Work, At Robert Morris Center, 401 South State Street, Suite 822 Chicago, IL 60605, United States

ARTICLE INFO

Article history:
Received 26 July 2016
Accepted 22 August 2016
Available online 28 August 2016

Keywords:
Group concept mapping
Measurement
Scale development
Psychometric testing

ABSTRACT

The past decade has seen an increase of measurement development research in social and health sciences that featured the use of concept mapping as a core technique. The purpose, application, and utility of concept mapping have varied across this emerging literature. Despite the variety of uses and range of outputs, little has been done to critically review how researchers have approached the application of concept mapping in the measurement development and evaluation process. This article focuses on a review of the current state of practice regarding the use of concept mapping as methodological tool in this process. We systematically reviewed 23 scale or measure development and evaluation studies, and detail the application of concept mapping in the context of traditional measurement development and psychometric testing processes. Although several limitations surfaced, we found several strengths in the contemporary application of the method. We determined concept mapping provides (a) a solid method for establishing content validity, (b) facilitates researcher decision-making, (c) insight into target population perspectives that are integrated a priori, and (d) a foundation for analytical and interpretative choices. Based on these results, we outline how concept mapping can be situated in the measurement development and evaluation processes for new instrumentation.

1. Introduction

Concept mapping is framed as an inclusive, participatory, collaborative, and inductive social science research process (Kane & Trochim, 2009). The methodology's flexibility is recognized as a strength, and the number of topics for which the method could be applied seems virtually limitless. It enables both detailed idea generation by stakeholders and higher-level conceptual representation. Although sophisticated multivariate analyses are employed, the results are visual and intuitive, thereby enhancing interpretation and use (Kane & Trochim, 2007; Trochim, 1989). Over its 25 year history, concept mapping has been used in an array of fields to develop theory, plan for programs and social interventions, evaluate social programs, and develop measures and scales (Kane & Trochim, 2009).

The foundation for the use of concept mapping in measurement was outlined in the early development and articulation of the method. Drawing from Campbell (1966); Campbell (1986) ideas about the natural coherence between observable patterns in both theory and reality, Trochim (1985) framed social research as a pattern matching exercise that involves correspondence between conceptual and operational domains. This provided the philosophical and epistemological basis for concept mapping as a technique for explicating a conceptual domain. Later writings concentrated on the methodological tenets of concept mapping as an integrated, mixed-methods approach that enabled groups to conceptualize an issue of relevance (Trochim & Linton, 1986; Trochim, 1989). The publication of a special issue in *Evaluation and Program Planning* in 1989 introduced concept mapping to the broader community of evaluators and researchers, and offered a practical and useful conceptualization tool for managing diverse perspectives and distributed group knowledge. Two early studies highlighted the application of concept mapping to measurement development (e.g., scales, measures, questionnaires). Galvin (1989) used the method to organize a stakeholder-produced conceptual framework from which an evaluation questionnaire was directly constructed. In generating additional content of relevance, vander Waal, Casparie, and Lako, 1996 used concept mapping to

* Corresponding author
E-mail address: srosas@conceptsystems.com (S.R. Rosas).

http://dx.doi.org/10.1016/j.evalprogplan.2016.08.016
0149-7189/

purposefully include representatives of the intended targets of the measure, and emphasized their contributions to the clarity and validity of the instrument. Although these studies lacked a complete description of the practical, step-wise application of concept mapping in the context of traditional scale development and psychometric testing procedures, they suggested the flexibility of the method to support and enhance measurement quality.

Since 2000, a body of measurement development research within the social, behavioral, and health sciences that includes concept mapping as a primary technique has emerged. Several aspects of the method have likely contributed to its presence in the literature. The generation of a large set of ideas, structuring of ideas based on judgments made about their interrelationships, graphical representations of scaled similarities among theoretical ideas, and identification of clustered sets of like items are some practical features that align concept mapping with general approaches for measurement development (Kane & Trochim, 2009). In their contemporary 8-stage mixed-methods framework for instrument development, Velozo et al. (2012) recommended concept mapping as a structured qualitative method for conceptualizing the construct(s) to be measured and developing representative items. They further emphasized the method's value in synthesizing literature findings, operationally defining constructs, and generating hypotheses about the scope and content of the scale. Previous research has demonstrated concept mapping to be a valid and reliable conceptualization approach in general (Rosas & Kane, 2012). However, little guidance or understanding is available on how concept mapping can and should be integrated in the measurement development process. Furthermore, despite the purported epistemological, methodological, and ontological value of concept mapping, little has been done to critically review how researchers have approached the application of the method in applied measurement development research.

To that end, we systematically reviewed the literature on concept mapping to identify where and how the method was applied in the context of measurement development and evaluation. In this review we examine the practice of integrating the concept mapping methodology into processes for establishing new measurement tools in accordance with generally accepted development and testing procedures based on established psychometric principles. From this examination we assess the current practice of using concept mapping in applied measurement development research, noting the strengths, limitations, and future directions for the field.

1.1. Scale and measurement development in social science research

To begin, it is useful to broadly outline the measurement development and psychometric evaluation process within social, behavior, and health sciences research. This multi-step process generally involves the (a) articulation of construct(s) of interest and their context, (b) specification of the response format and selection of the initial items, (c) collection of data from a set of target respondents, and (d) examination of the psychometric properties and determination of quality (DeVellis, 2011; Furr, 2011; Simms, 2008).

Formal development activities are conducted to protect against two types of error: measuring less than the proposed construct (i.e., construct underrepresentation) and measuring more than the proposed construct (i.e., construct irrelevant invariance). Rigor in the process of conceptualization and definition is required to avoid the first type of error. Establishing content validity – the minimum psychometric requirement for measurement adequacy – relies on sufficiently capturing the specific domain of interest, while simultaneously containing no extraneous content (Netemeyer et al., 2003; Schriesheim, Powers, Scandura, Gardiner, & Lankau, 1993). Rigor in psychometric analysis is required to avoid the second type of error. Reliability and validity are fundamental facets of psychometric quality and researchers strive to provide evidence regarding the nature and strength of these characteristics (Furr, 2011). Psychometric quality is further demonstrated in the assessment of the instrument's performance in the sample being studied through results that truly reflect the hypothetical construct(s) it purports to measure (DeVellis, 2011). It is within this ongoing, iterative process, that use of concept mapping as a core method in measurement development and evaluation is reviewed.

2. Method

2.1. Review sample selection

Our review began with a literature search to identify a sample of published studies where concept mapping was employed as a principal method in the measurement development process. Due to the range of fields where concept mapping has been used, we determined a broad search was warranted using several highly-cited publications as sources. We identified three seminal publications, Trochim and Linton (1986), Trochim (1989), and Kane and Trochim (2007) as the most frequently referenced source publications for the concept mapping methodology. Using these three sources as the point of reference, a Google Scholar search returned lists of 152, 894 and 323 other works (i.e., published literature, grey literature, reports, etc.) citing these publications, respectively. We further narrowed the three lists by filtering each through the following search string: “scale development OR measurement OR content validity OR psychometric testing”. This filtering step returned 145, 434, and 99 works, respectively. From these results, we then applied specific criteria for inclusion into our review set. First, the work had to be a published study in a peer-reviewed journal. Second, the study either (a) outlined the development of a conceptual measurement model/framework using concept mapping, or (b) referenced the development of a conceptual measurement model/framework using concept mapping. Third, a new measurement tool was created and psychometrically evaluated, either within the same study or in a subsequent publication. Several studies initially identified specifically mention the use of concept mapping and the construction of a scale based on the results of the process (cf. Armstrong & Steffen, 2009; Iris, DeBacker, Benner, Hammerman, & Ridings, 2012; Shorkey, Windsor, & Spence, 2008; Shorkey, Windsor, & Spence, 2009). However, this group of studies lacked a complete examination of psychometric properties and a separate validation study of the scale could not be found elsewhere in the literature. Thus, they were not included.

2.2. Review sample

In applying the aforementioned criteria, we identified 23 published studies between 2001 and 2014. The use of concept mapping in measurement development and evaluation appears to be a fairly recent practice, with all identified studies occurring after 2000. Two studies identified in the initial query were not included in the review, but are noteworthy. These studies were unique in the application of concept mapping for examining and improving *existing* scales. White and Farrell (2001) used concept mapping with a small sample of experts to revise an original conceptual model and conduct an analysis of secondary data using confirmatory factor analytic techniques to determine the most parsimonious structural representation of items. Sepúlveda Carrillo, Meneses Báez, and Goldenberg, 2014 used concept mapping post-hoc to evaluate the conceptual structure and item sequence of a

previously developed, reviewed, and pilot-tested questionnaire, analyzing the sorting and rating information supplied by 16 experts. In both cases, researchers claim the revised scales and measures are improved versions in terms of comprehensiveness with fewer items, stronger psychometric characteristics, and better alignment with contemporary theoretical constructs found in their respective fields.

2.3. Scope of review

Our review involved examining each study in terms of how they describe elements of the concept mapping process, as outlined by Trochim and colleagues. Given the mixed-method nature of concept mapping, our review was facilitated by developing a database of common elements and summarizing well-known qualitative and quantitative characteristics. In this review, we examined the (a) purpose, (b) focus prompt, (c) stakeholders and perspectives, (d) idea generation, (e) structuring – sorting and rating, and (f) interpretation elements of the concept mapping studies. In addition, we examined activities where concept mapping results were used to inform key decisions in the measurement development process, such as item refinement and selection. Finally, we reviewed the psychometric evaluation process of each study, noting the types of analytical tools employed during the psychometric testing procedures. As with concept mapping, we summarized common elements related to sampling, statistical methods, and general psychometric results. The purpose here was to identify patterns in the psychometric evaluation process, the relationship to concept mapping, and whether differences were present in the way conventional analytic approaches were used in the concept mapping studies. We assume that a full evaluation of the appropriate use of the analytical methods occurred during the peer-review process. Thus, a critique of the psychometric methods or a pooled analysis of reliability and validity estimates is beyond this review.

3. Results

This review focused on the 23 measurement development studies listed in Table 1. Two subsets are included in this final sample. In one set (n = 15), concept mapping, scale and measure construction, and psychometric testing and evaluation processes were reported within a single manuscript. In the other set (n = 8), the aforementioned processes were published separately. In these paired publications, the first described the use of concept mapping to generate conceptual structure of the content for a particular topic of interest. These conceptualization studies clearly outline the focus of the concept mapping inquiry, describe the steps in the concept mapping process, detail the results and interpretive findings, and indicate the anticipated utilization of the results, clearly emphasizing the future development and testing of a scale or measure. The second publication of the pair detailed the instrument construction steps and formal psychometric assessment. Although varied and described in much greater detail in the separately published studies, researchers routinely reported on the general procedural steps for conducting concept mapping found in the seminal literature on the method (Kane & Trochim, 2007; Trochim & Linton, 1986; Trochim, 1989).

3.1. Purpose

The majority of the studies in this review were primarily found in the areas of health care (e.g., patient care, health education, quality of life) and social welfare (e.g., elder abuse and exploitation, mental health services). The purpose outlined by researchers for undertaking the measurement development study was categorized into five major groups. First, a set of studies focused on the creation of an instrument to be used in the evaluation of an intervention (Ciciriello, Buchbinder, Osborne, & Wicks, 2014; Osborne, Elsworth, & Whitfield, 2007; Rosas & Camphausen, 2007). Second, several studies emphasized the purpose of the research was in response to a need to better conceptualize and measure complex phenomenon (Battterham et al., 2002; Behfar, Mannix, Peterson, & Trochim, 2011; Conrad, Iris, Ridings, Langley, & Anetzberger, 2011; Conrad, Iris, Ridings, Langley, & Wilber, 2010; Iris, Conrad, & Ridings, 2014; Jordan et al., 2013; Wolfinbarger & Gilly, 2003). Third, a set of studies focused on developing more comprehensive instruments that addressed the limitations (e.g., coverage, content, scope) found in previous efforts (Osborne, Batterham, Elsworth, Hawkins, & Buchbinder, 2013; Rosas, Behar, & Hydacker, 2014; Van Haitsma et al., 2012; van Nieuwenhuizen, Schene, Koeter, & Huxley, 2001). Fourth, some studies sought to address the need for and lack of a valid and reliable assessment for a specific condition or experience (Butler, Budman, Fernandez, & Jamison, 2004; Butler et al., 2007; Corcoran, 2005; Luke, Calhoun, Robichaux, Elliott, & Moreland-Russell, 2014; Osborne et al., 2011; Wallace, Wexler, Miser, McDougle, & Haddox, 2013). Finally, a small set of studies focused on the development of an instrument to capture individual perspectives related to quality of care – a newly emerging area for assessing the impact of services upon recipients (Beijersbergen, Asmoredjo, Christians, & Wolf, 2014; de Kok et al., 2007; van der Eijk et al., 2001).

3.2. Focus prompt

Concept mapping employs the use of a single statement, often referred to as a focus prompt, that participants respond as they generate ideas. Typically developed by researchers and worded in a way to provide specific instruction, the focus prompt is the trigger for the idea generation process and establishes the boundaries of the conceptualization (Kane & Trochim, 2007). Functionally, the prompt serves as the reference for participants as they identify content of a particular topic, which in the case of the studies reviewed here, served as the source material for measurement tool construction.

The focus prompts used in these studies varied slightly despite their common purpose in setting the parameters for content generation that informed the development of a theoretical measurement pattern. Several studies framed the focus prompt in a way to direct participants to think generally on the topic and used a similarly worded phrase, “Thinking as broadly as possible, generate statements that . . . ” (Southern, Young, Dunt, Appleby, & Batterham, 2002; Osborne et al., 2011); Osborne et al., 2007) or ‘Thinking as broadly as possible, please list specific . . . ’ (Wallace et al., 2013). The focus prompts of other studies were more specific in terms of the parameters of the content desired for the type of instrument anticipated. These prompts were framed towards identifying outcomes or end results (“Generate statements which describe the specific benefits that family members engaged in the family support program should experience” (Rosas & Camphausen, 2007)), risks or concerns (‘Please list indicators, or risk factors, of potential problems for opioids in patients considered for opioid therapy’ (Butler et al., 2004)), or symptoms (" . . . [list] symptoms typical of youth in the juvenile justice system” (Corcoran, 2005)). Still, others were more purposeful and direct in framing the focus prompt as measurement development activity, for example, “Generate a short statement that describes an item that should be included in an elder self-neglect measure” (Iris, Ridings, & Conrad, 2010). Although it is plausible that the prompts could be framed differently and still yield the desired content, the focus prompts in the studies in this review appeared align with the study

Table 1
Scales and questionnaires developed using concept mapping and psychometrically evaluated, listed alphabetically.

Scale or Questionnaire	Type	Focus	Concept Mapping		Psychometric Testing		Sources(s)
			N	Description	N	Description	
Consumer Quality Index for Shelter and Community Care Services (CQI-SCCS)	S	Experiences of homeless adult and youth and abused women with shelter and community care services.	B: 22 S/R: 161	Clients and executive staff of homeless care institutions.	I: 762 F:118	Clients of organizations providing services to homeless adults and youth, and abused women	Beijersbergen et al. (2014)
Current Opioid Misuse Measure (COMM)	SI	Aberrant medication-related behaviors of chronic pain patients prescribed opioid therapy	B: 26 S/R: 26	Professionals from the International Pain and Chemical Dependency Listserv	I: 227 F: 55	Patients taking opioids for chronic non-cancer pain	Butler et al. (2007)
Elder Self-Neglect Assessment (ESNA)	O	Self-neglect behaviors of older adults	B: 20 S/R: 50	Services program supervisors, Geriatricians; Local policy analysts and program planners; Elder law practitioners, University-based researchers, Case managers and supervisors, Elder abuse investigators, Social workers	I: 215	Clients of 11 aging services agencies	Iris et al. (2010, 2014)
GP Integration Index	S	Integration behaviors of physicians between primary and secondary care sectors	B: 173[a] S/R: 173[a]	Consumer representatives; Hospital administrators; Specialist doctors; Community service providers; Nurses; Allied health providers, General practitioner groups	I: 501 F: 151	National (Australian) probability sample of General practitioners – Solo, Hospital, and Medical-center Based	Southern et al. (2002); Battterham et al. (2002)
Health Education Impact Questionnaire (heiQ)	S	Outcomes of health education programming	B: 17[a] S/R: 17[a]	Patient education participants with chronic illnesses' Program managers; Health professionals, Course leaders, Policymakers	I: 591 F: 598	Consumers of patent education programs and hospital outpatients	Osborne et al. (2007)
Health Literacy Management Scale (HeLMS)	S	Individual's capacity to seek, understand, and use health information	B: 15[a] S/R: 15[a]	Participants of education programs with and without chronic health conditions	I: 333 F: 350	Patents in chronic disease self-management programs and emergency room attendees	Jordan et al. (2013)
Health Literacy Questionnaire (HLQ)	S	Individual's motivation and ability to seek, gain, understand and use information to promote health	B: 15[a] S/R: 15[a]	Participants of education programs with and without chronic health conditions	I: 634 F: 405	Patients reviving services from emergency room, private specialist clinic and home and community health services organization	Osborne et al. (2013)
Individual Parent Strengths and Capabilities Questionnaire (IPSCQ)	S	Behavioral strengths and capabilities of parents in family support programming	B: 14 S/R: 14	Family support program mangers and staff	I: 268	Parents and primary caregivers in family support programs	Rosas and Camphausen (2007)
Influenza Intensity and Impact Questionnaire (FluiiQ™)	S	Perceived effects of influenza infection	B: 16 S/R: 16	Patients with confirmed influenza	I: 311	Patients with influenza-like illness across 25 sites (US)	Osborne et al. (2011)
Lancashire Quality of Life Profile (LQoLP)	SI	Quality of life	B: 29 S/R: 29	Mental health patients; Relatives of psychiatric patients; Professional caregivers	I: 518 F: 31	Long-term mental health patients	van Nieuwenhuizen et al. (2001)
Methotrexate in Rheumatoid Arthritis Knowledge Test (MiRAK)	S	Patient knowledge about Methotrexate treatment for Rheumatoid Arthritis	B: 24 S/R: 24	Patients with Rheumatoid Arthritis who received treatment with Methotrexate	I: 169 F: 131	Patients with Rheumatoid Arthritis treated with Methotrexate	Ciciriello et al. (2014)
Older Adult Financial Exploitation Measure (OAFEM)	SI	Financial exploitation of older adults by caregivers	B: 16 S/R: 16	Local public and non-profit service provider and agency representatives; Nationally recognized elder abuse scholars	I: 227	Clients of 7 adult protective services agencies	Conrad, Iris, Ridings, Rosen et al. (2011) Conrad et al. (2011)
Older Adult Psychological Abuse Measure (OAPAM)	SI	Psychological abuse of older adults by caregivers	B: 16 S/R: 16	Local public and non-profit service provider and agency	I: 226	Clients of 7 adult protective services agencies	Conrad, Ridings et al. (2011) Conrad et al. (2010)

				representatives; Nationally recognized elder abuse scholars			
Online etail quality (eTailQ)	S	Quality of the online retailing experience	B: 64[b] S/R: 90	B: Graduate students, Faculty, Online shopping public S/R: Graduate and undergraduate students who made online purchases	I: 1013	Randomly selected adults over 18 years of age	Wolfinbarger and Gilly (2003)
Oregon Mental Health Referral Checklist (OMHRC)	S O P	Mental health needs of youth in the juvenile justice system	B: 15 S/R: 15	Program administrators and providers in mental health and juvenile justice	S, I: 83 O, I: 146 P, I: 52	Adjudicated and incarcerated youth; Parents	Corcoran (2005)
Patient Opioid Education Measure (POEM)	SI	Patient knowledge and expectations regarding chronic opioid use	B: 14 S/R: 37	Primary care physicians; Pain/addiction specialists; Clinical psychologist; Researcher; Pharmacists; Patient education librarian	I: 83	Patients taking opioid medication for chronic non-cancer pain	Wallace et al. (2013)
Preference for Everyday Living Inventory (PELI)	S	Daily life preferences for older adults in person-centered care delivery	B: UNK S/R: 20	B: Researchers and focus group participants S/R: Older gerontologists with established records of research and service	I:528	Home health agency clients	Carpenter et al. (2000); Van Haitsma et al. (2012)
Process Conflict Scale (PCS)	S	Management of logistics and coordination in accomplishing group tasks	B: 225 S/R: 20	B: Cohort of MBA students in business school working in teams S/R: MBA student enrolled in different business school	I: 182 F: 885	Management graduate students working in teams	Behfar et al. (2011)
Program Sustainability Assessment Tool (PSAT)	S	Capacity for public health programs to sustain programs, policies, and activities	B:106 S/R: 39	Experts in public health and program sustainability representing scientific institutions, funding and advisory agencies, and state and community programs	I: 592	Program managers and staff of 252 public health programs	Schell et al. (2013); Luke et al. (2014)
Quality of Health Care Breast Cancer (QUOTE-BC)	S	Patient perceptions of quality of breast cancer health care	B: 72 S/R: 67	Patients from different hospitals implementing specific breast cancer treatment regimens	I: 276	Breast cancer patients in 5 hospitals experiencing surgery within the previous 3–15 months	de Kok et al. (2007, 2010)
Quality of Health Care in Inflammatory Bowel Disease (QUOTE-IBD)	S	Quality of care for patients with Inflammatory Bowel Disease	B:267 S/R: 30	B: Patients with Inflammatory Bowel Disease from 6 countries S/R: Patients with Inflammatory Bowel Disease from the Netherlands	I: 162 F: 118	Patients with Inflammatory Bowel Disease from the Netherlands	van der Eijk et al. (2001)
Screener and Opioid Assessment for Patients with Pain (SOAPP)	S	Potential risk of opioid abuse for chronic pain patients considered for long-term opioid therapy	B: 26 S/R: 39	B: Pain specialists, Primary care providers, Nurses, Medical center support staff S/R: Professionals from the International Pain and Chemical Dependency Listserv	I: 116 F: 95	Patients taken or considered for long-term opioid medication regiment	Butler et al. (2004)
System of Care Readiness and Implementation Measurement Scale (SOC-RIMS)	S	Community-level elements necessary to develop a system of care for children's mental health service	B: 135 S/R: 36	Stakeholders from grant-funded communities, Nationally recognized experts, consultants, trainers, and leaders	I: 506	System of care stakeholders from 24 sites	Behar and Hydaker (2009), Rosas et al. (2014)

Note: B = Brainstorming/Idea Generation; S/R = Sorting/Rating; I = Initial sample; F = Follow-up sample; P = Parent-report; O=Observation; S = Self-report; SI = Structured interview; UNK = Unknown.
Note: Single study measurement publications (n − 15) include one paper in the Source(s) column; Paired study measurement publications (n = 8) include both papers in the Source(s) column.
[a] Multiple concept mapping procedures and outputs.
[b] Participants in traditional focus groups from which ideas were extracted.

purpose and the need to generate an expansive set of statements in the domain of interest.

3.3. Stakeholders and perspectives

In order to ensure a broad range of ideas on the topic, the identification and inclusion of participants whose knowledge or opinion meaningfully contributes to the resulting framework is critical. As a participatory method, concept mapping requires thoughtful consideration of the match between participant insight and perspective and the focus of the study. The adequacy of the content domain is bounded by the source of input, and therefore it is dependent upon the correspondence between the aim of the conceptualization and the participants included in the process.

Across the 23 studies, the number of participants in the core concept mapping processes (i.e. brainstorming, sorting, and rating) were consistently reported. As shown in Table 1, the number of brainstorming and sorting/rating participants were accounted for in each study. In these studies, the sorting and rating participants were the same, although it is common for sorting and rating participants to constitute different groups within the same study (Kane & Trochim, 2007; Rosas & Kane, 2012). Including those studies where separate group concept mapping processes (i.e. multiple concept mapping sessions and results) were conducted (cf. Osborne et al., 2007; Southern et al., 2002) the average number of participants in the idea generation step was 37.91 (SD = 59.07; Range = 7–267) and the average number of sorting and rating participants was 26.80 (SD = 17.65; Range = 7–161). The average number of sorters is consistent with previous *meta*-analytic studies on the method by Trochim (M = 14.62; 1993) and Rosas and Kane (M = 24.63; 2012). The average number of raters, although higher than what Trochim (1993) found in his early *meta*-analytic study (M = 13.94), is substantially lower than estimates found in a more recent *meta*-analytic study (M = 81.77; Rosas & Kane, 2012).

Descriptively, the sample of concept mapping participants appears to be directly linked to the focus of the measurement development process outlined by the researchers. Three types of participant groups were identified across the 23 studies (see Description column in Table 1). First, a number of the studies (n = 8) only included representatives of the target population for whom the instrument was intended. Second, a larger set of studies (n = 10) reported that concept mapping participants were experts in the content area or topic, either as scholars or direct service professionals. A third, smaller set of studies (n = 5) blended both expert and target perspectives, depending upon the focus of the measurement tool. Although studies were inconsistent with reporting the mechanisms for concept mapping participant recruitment and management of input, they were fairly clear in the intent and purpose for participant inclusion, emphasizing diversity of perspective in relation to role, experience or situation.

3.4. Idea generation

Articulation of the content domain, from which items may be selected, is arguably the most critical part of developing a sound measurement tool. In concept mapping, the generation of content is typically conducted through a brainstorming procedure with participants (Osborn, 1957), although other sources of information such as interviews, reports, and literature can and have been used (Kane & Trochim, 2007). The primary consideration in the content generation step is the production of a comprehensive set of statements that ideally represents the broadest conceptual domain of the topic of interest.

The open brainstorming process in response to a focus prompt, as outlined by Trochim and colleagues, was the most frequent method for generating content in this sample (n = 17). However, several researchers raised concerns about the comprehensiveness of the content generated through the brainstorming process and included additional methods for expanding the content (Conrad et al., 2010; Conrad et al., 2011; Schell et al., 2013). Several studies used findings from literature reviews and syntheses to supplement brainstormed lists. Only three of the studies did not use brainstorming, instead electing to extract statements from content derived through conventional focus group methods (Wolfinbarger & Gilly, 2003; de Kok et al., 2007), in addition to literature review results (Carpenter, Van Haitsma, Ruckdeschel, & Lawton, 2000). In a few cases, the idea generation step of concept mapping process was enhanced with existing scale items from other measures (Wolfinbarger & Gilly, 2003; van Nieuwenhuizen et al., 2001).

Often, the idea generation methods used in these studies yielded many more items that what could be reasonably managed in the sorting and rating process. The majority of the scale development concept mapping studies included some form of idea syntheses process, which is an effort to reduce the number of items originally generated to an appropriately manageable set for inclusion in subsequent steps. Several studies (n = 12) reported generating several hundred statements, with as many as 400–500. Kane and Trochim (2007) recommend fewer than 100 items are optimal and meta-analyses on concept mapping have shown that on average concept maps include about 84 (Trochim, 1993) to a little more than 93 statements (Rosas & Kane, 2012). In this sample of studies, the average number of statements included in the final concept map (M = 98.23; SD = 39.15; Range = 54–235) was consistent with previous meta-analyses of concept maps. While some researchers implemented a supplemental process to the brainstorming step that worked to reduce the redundancy of items and those off topic (Conrad, Iris, Ridings, Rosen et al., 2011; Conrad, Ridings et al., 2011; Iris et al., 2014), the majority of the studies reviewed here did not routinely report the methods used or the guidelines employed to reduce the original set of statements from the content generation process to its final set for the next step of the concept mapping process.

3.5. Structuring – sorting and rating

The structuring step in concept mapping involves two participant data collection tasks: sorting and rating. In this step, each concept mapping participant arranges the set of ideas into groups based on perceived similarity, and then rates each idea based on one or more scales. To produce the concept map, non-metric two-dimensional multidimensional scaling (MDS) of the sort data (Davison, 1983; Kruskal & Wish, 1978) and a hierarchical cluster analysis (HCA) of the MDS coordinates (Everitt, 1980) are conducted.

Studies in this sample consistently reported on the details of the structuring step in the concept mapping process. Nearly all (n = 22) described the process where participants individually categorize content based on similarity using an unstructured sorting procedure as outlined by Trochim and colleagues. All studies indicated the use of multidimensional scaling (MDS) analysis of the sorting data, although the level of detail regarding MDS and its characteristics (similarity) varied. All studies reported the use of cluster analysis, however there was variation in the degree to which specifics regarding the type (hierarchical, agglomerative), source of data (x-y coordinates from the MDS analysis), and algorithm (Ward's) outlined by Trochim and colleagues were indicated. Few (n = 4) of the studies in this sample reported the stress value of the MDS procedure. Stress is a statistical value routinely generated in MDS analyses and reflects the goodness of fit between the final representation and the original similarity matrix used as input (Kruskal, 1964). The stress value has been

described as an indicator of internal representational validity (Rosas & Kane, 2012) and reflects the degree to which the conceptualized model (i.e., the concept map) reflects the judgments made by participants as a function of the sorting procedure. Despite its absence, the majority of the studies ($n = 16$) did display the concept map as a means for representing the final conceptual model produced by the concept mapping procedure.

More than half ($n = 16$) of the studies in this sample also reported at least one rating of the statements in the final set. The rating dimension of importance was the most frequently used ($n = 11$). Other ratings included severity, impact, immediate need, duration, modifiability, consistency, and difficultly implementing. Only a handful of studies reported separately the results of the ratings analysis, usually in the form of average item and average cluster ratings in table format. However, several studies integrated the ratings results with the representation of the concept map in the form of a "cluster rating map" that shows the average of the cluster as stacked layers of the polygon-shaped cluster perimeter (Butler et al., 2004, 2007; Conrad, Iris, Ridings, Rosen et al., 2011; Conrad, Ridings et al., 2011; Corcoran, 2005; Iris et al., 2010; Rosas & Camphausen, 2007). Furthermore, a few studies displayed the average cluster ratings in a "ladder graph" to visually compare and correlate the values between two groups across the cluster arrangement (Butler et al., 2004, 2007; Wallace et al., 2013) and one used a bivariate scatter plot to contrast the item averages from two different ratings (Behar & Hydaker, 2009).

3.6. Interpretation

Interpretation of the concept map is a formal step in the concept mapping process and the rationale for engaging a variety of stakeholders in this step is articulated by Kane and Trochim (2007). Relational nuances found in the map, cluster selection, and cluster label finalization are typically processed during this step, and decisions have implications for the theoretical pattern that surfaces. As a rule, a group collaboratively interprets the maps that result from the analyses through a process of review, deduction, and labeling of the clusters in a substantively meaningful way.

In this sample, interpretation of the concept maps was primarily managed by the research team. The majority of studies ($n = 14$) indicated that only the researchers were involved in the interpretation step. However, a number ($n = 9$) of studies included experts in addition to the research team, and in a few cases included some the participants of previous steps (i.e., brainstorming, structuring). The rationale for the make-up of the interpretation group was not typically offered, except in those cases when previously engaged participants were involved, emphasizing their role in enhancing the credibility, trustworthiness, accuracy, transferability and coherence of the conceptual model (Conrad, Iris, Ridings, Rosen et al., 2011; Conrad, Ridings et al., 2011; Rosas & Camphausen, 2007; Southern et al., 2002). The interpretation step described in these studies also included a review of the statements, as it pertained to the construction of measurement instruments. This is somewhat unique from typical concept mapping studies, as there is an indicated need to consider and prepare concept mapping results for utilization in the next steps of the measurement development process.

3.7. Measurement tool item selection

Ensuring adequate coverage of the domain, while simultaneously achieving parsimony, is a critical balance in the pursuit of content and construct validity (Cronbach & Meehl, 1955). Decisions regarding length are significant, as shorter scales and measures help minimize respondent fatigue and response biases (Hinkin, 1998; Streiner, Norman, & Cairney, 2014). However, too few items can seriously affect content and construct validity, internal consistency and test-retest reliability (Converse & Presser, 1986). Despite the lack of guidance and recommendations on item reduction and refinement for concept mapping, such a process was a key to instrument construction with implications for psychometric testing.

Overall, there was variation in how researchers linked the results of the concept mapping portion of the process, with the instrument construction activity. More than half of the studies ($n = 16$) reported some form of item reduction following the finalization of the conceptual structure as represented by the concept map. Only a few studies used the full concept mapping derived set of statements as the instrument items, converting each concept mapping statement to an instrument item (Battterham et al., 2002; Rosas et al., 2014). Others added items based on expert consensus post-concept map (Conrad et al., 2010; Iris et al., 2014; Osborne et al., 2013), although this increase was minor (10 or fewer). Some researchers used qualitative tools and methods, such as literature reviews, logic models, consultations, and interviews to confirm and guide decision-making on which items to retain or add (Conrad et al., 2011; Conrad et al., 2010; Jordan et al., 2013; Osborne et al., 2007, 2011). In the cases where concept mapping ratings results were used to facilitate decisions regarding item reduction, the use of simple criteria were described, such as selecting items with an average rating above a specific level (Corcoran, 2005; Wallace et al., 2013). However, some investigators employed more sophisticated criteria for item selection. For example, Rosas and Camphausen (2007) selected items with an item-total correlation value above 0.70 and van der Eijk et al. (2001) retained those items that loaded on a single factor in factor analyses of the ratings. Other studies where ratings results were used to guide item reduction lacked specificity, simply indicating selection of the highest rated items in each cluster (Butler et al., 2007; Wolfinbarger & Gilly, 2003) and "researcher-determined" criteria applied to mean scores (de Kok et al., 2010). To further refine the items included in the initial scales and measures, a few studies sought to ensure understandability and readability using cognitive interviewing (Jordan et al., 2013; Osborne et al., 2011; Osborne et al., 2013) or calculating Lexile scores (Wallace et al., 2013).

Irrespective of the methods used to condense the number of statements from the concept map to the initial scale items, the average reduction was sizeable. Across 16 studies where a reduction was identified, the average number of items eliminated from the final set of concept mapping statements was 55.52 ($SD = 54.46$; Range = 0–225), with sometimes more than half of the statements eliminated. The initial versions of the instruments were substantially smaller than the original set of statements compiled during the idea generation process. Across the 23 studies presented in Table 1, the average number of items in the initial scale or measures was 57.43 ($SD = 28.06$) items and ranged from 10 to 114. As shown in Table 1, the majority of the scales and measures developed using concept mapping were designed as to be self-report ($n = 16$). A few were designed as structured interviews ($n = 5$) and one was designed as an observation, intended to be facilitated by an experienced professional in the context of services provision. Of note, one study reported on the development of parallel versions of three forms: a self-report instrument, an observational checklist completed by professionals, and a parent-reported instrument (Corcoran, 2005).

3.8. Psychometric testing and evaluation

To evaluate the overall quality of the scale or measure and judge the performance of individual items, the draft instrument is administered to an appropriately large and representative sample.

By attending to dimensionality, reliability, and validity of a scale or measure researchers enhance the possibility that the instrument will be useful and informative. Although all of the studies reviewed here employed concept mapping to help inform the development process we observed diversity in the statistical methods used to determine the psychometric quality of the scales or measures. Table 2 displays the variety of analytical techniques used to assess validity and reliability across the set of scale and questionnaire development studies. Nearly all ($n = 18$) of the studies employed multiple analytic techniques to assess multiple forms of validity of the scale or measure.

A clear description of the sample, the sampling technique, response rates, and the questionnaire administration process was provided in nearly every study included in this review. Descriptively, the psychometric testing samples reported in the studies appeared to correspond with the focus of the scale or measure, and there was a substantial concurrence between the psychometric testing sample and the concept mapping participants (see Table 1). While all studies reported information on an initial sample of target participants, several ($n = 11$) also included a description and purpose for a follow-up sample in order to conduct additional psychometric tests. As shown in Table 1, the initial psychometric testing samples were sizable, averaging 348.84 ($SD = 242.86$; Range = 52–1013) participants across studies in the sample, despite a few studies with samples below 100 participants. The follow-up samples in the 11 studies were lower, averaging 267 ($SD = 269.07$; Range = 31–885) participants. Some studies ($n = 5$) reported employing randomized selection of participants, whereas the majority ($n = 18$) relied upon non-random means for selection of participants, such as convenience samples of individuals within a specific setting. Sample size decisions appeared to be dictated by the choice of analytical approach, with the largest groups of study participants reported in those studies where factor analytic techniques were used.

Factor analysis is the most commonly used analytic technique for data reduction and refining constructs (Floyd & Widaman, 1995) and as shown in Table 2 was the most frequently used analysis in the reviewed studies. Eighteen studies used exploratory or confirmatory factor analysis, some in combination to examine the stability of the factor structure and provide information to facilitate the refinement of a new measure. Nearly all of the studies employing these techniques presented clear and thorough descriptions of factor analytical techniques and results as advocated by Tabachnick and Fidell (2001). Both exploratory and confirmatory factor analysis have been shown to be particularly susceptible to sample size effects. Sample sizes appeared to be appropriate for the type of analysis, with sizes ranging from 311 to 1013 for confirmatory techniques and 182–726 for exploratory techniques. Sample sizes were consistent with research suggesting roughly 150 observations for exploratory factor analysis (Worthington & Whittaker, 2006) and 200 observations for confirmatory factor analysis (Kline, 2006) are sufficient to obtain an accurate solution, provided item intercorrelations are reasonably strong.

Seven of the studies employed item response theory techniques, specifically Rasch analysis for the examination or dimensionality, validity and reliability. These studies presented clear and thorough descriptions of Rasch analysis techniques and results as advocated by Embretson and Reise (2000). Again, sample sizes appeared to be appropriate for this type of analysis although more narrow that those found in factor analytic samples, with sizes ranging from 169 to 227. These sample sizes were consistent with research suggesting that Rasch analysis is most efficient and accurate with samples larger than 100 participants (Chen et al., 2014). Of note, a few studies employed combinations of classical test theory techniques (i.e. factor analysis) and item response theory techniques (i.e. Rasch analysis) in the same study (Osborne et al., 2007, 2011, 2013).

In terms of other forms of validity, studies in this review primarily focused on (a) examining the congruence of the target construct(s) and the relative position to other similar and dissimilar constructs in the nomological net, and (b) assessing responses of groups who would be expected to differ on the instrument. Differences in the scores between known groups were assessed in 16 studies employing a wide variety of analyses (see Table 2). Convergent or divergent validity was assessed in 11 of the studies, primarily through correlation analysis. With respect to criterion-related validity, most of the studies in the sample focused on specific relationships that were theoretically justified in the introduction and literature review section of the study. Hypothesized relationships were usually examined and confirmed using

Table 2
Psychometric methods used across studies.

Property	Type	Analytic Technique	Number of Studies
Validity	Structural	Confirmatory Factor Analysis (CFA)	9
		Exploratory Factor Analysis (EFA)	5
		Principal Components Analysis (PCA)	4
		Rasch Analysis	7
	Convergent/Divergent	Correlation Analysis (Spearman Rank; Pearson)	11
	Known groups	Regression Analysis (Logistic, Linear)	2
		Correlation Analysis (Spearman Rank; Pearson)	5
		Analysis of Variance (ANOVA)	3
		Analysis of Covariance (ANCOVA)	2
		Mean Differences (e.g., T-test)	3
		Non-parametric group analysis (Mann-Whitney)	1
	Predictive	ROC Curve Analysis	2
Reliability	Internal consistency	Cronbach's Alpha	17
		Kuder-Richardson 20	1
		Rasch person reliability	4
	Temporal Stability (Test-Retest)	Cohen's Kappa	2
		Intraclass Correlation Coefficient (ICC)	6
		Pearson Correlation	2
	Parallel Forms	Coefficient of Equivalence	1

either correlation or regression analysis, and in four studies, using structural modeling.

In terms of reliability, internal consistency was the form most often assessed using Cronbach's Alpha, with 17 studies using this analysis to examine reliability. Concurrent with the use of Rasch analytic techniques described above, four studies examined Rasch person reliability, a measure analogous to Cronbach's alpha to estimate internal consistency. Internal consistency estimates were calculated for 77 components across the 23 scales or measures the median value of the estimates was 0.87 (Range = 0.15–0.97). Only one study produced poor internal consistency estimates across components. A number of studies examined the temporal stability of the scale or measure and the median value of the estimates was 0.79 (Range = 0.62–0.96).

Following formal psychometric testing, several of the scales or measures were further refined through the deletion of items not meeting a priori criteria. The rationale for the retention and deletion of items was clearly linked to the empirical results and supported theoretically by the researchers. Overall, the criteria used by investigators were clearly described and well-supported by generally accepted parameters for model and item fit, based on the measurement model used. Across the studies the average reduction in the number of items from the initial version of the scale or measure subjected to testing, to the final version was 22 (*SD* = 18.34; Range = 0–52) items. The average number of items in the final versions of the instruments was 36.37 (*SD* = 17.63; Range = 10–68).

3.9. Concept mapping – measurement instrument construction linkage

Psychometric analyses occasionally lead researchers to re-conceptualize the nature of the construct(s) that compose measurement tools. Upon revision, the instrument is again evaluated in terms of its conceptual and psychometric properties. This iterative process of creation, testing, revision, and re-testing is expected to result in scale or measure with good psychometric quality and clear conceptual meaning. Thus, examining the results relative to the initial conceptual model is important to understanding the relationship between the theoretical and observed domains.

Across this set of studies, the extent to which the psychometrically evaluated instrument was examined relative to the original concept map was minor. A large proportion of the studies made little to no reference regarding an explicit and purposeful linkage between the conceptualized model (i.e., the concept map) and the measurement model derived from the psychometric methods, but rather placed strong emphasis on individual items developed. Nonetheless, in some of the studies the researchers were intentional in operationalizing this linkage, albeit in a limited way. For example, Rosas and Camphausen (2007), Rosas et al. (2014), Van Haitsma et al. (2012), and Butler et al. (2007), examined the distribution of the final scale items retained as a result the psychometric analysis in relation to the original concept map. In these studies the researchers reported the proportion of items within each cluster that ended up being retained and assessed how the empirical model comports with the original conceptual model. Other studies reported the results of the psychometric analyses in relation to the conceptualized structure. In these cases, the psychometrically-derived estimates were reported according to the original structure of clusters and items and the validity and reliability of the resulting instrument was evaluated (Conrad et al., 2011; Conrad, et al., 2010; Luke et al., 2014). Finally, Wolfinbarger and Gilly (2003) and Behfar et al. (2011) identified and isolated constructs in the final measurement model within a larger semantic net of content originally identified in the concept mapping process. In these cases, researchers described the empirically supported model in relation to other constructs differentiated in the conceptualization process.

4. Discussion

According to McGrath (2005), the conceptual complexity of constructs and their associated measures have played an important role in the failure to develop more accurate measurement systems. The studies included in this review integrated concept mapping in a variety of ways to establish a clear conceptual measurement framework. The new instruments produced in these studies demonstrated adequate to strong psychometric properties through the use of multiple analytical methods following conventional practices. Based on this review of how concept mapping is used in applied measurement development research, a number of conclusions as to its application in this context were reached. Despite our observations, the direct effect of concept mapping upon the psychometric results found in these studies is unknown and we can only speculate on its impact. In general, the field has not yet systematically addressed the question of the degree to which qualitative methods improve the measurement qualities of a scale (Velozo, Seel, Magasi, Heinemann, & Romero, 2012). To do so would require a highly-controlled design that had a clearly established comparator method, a scenario that seems implausible and unlikely. While this question remains unanswered, structured mixed-methods that work to assure content validity seem not only necessary and often neglected, but relatively simple to accomplish. Below we discuss the insights regarding the strengths and limitations from our review separately.

4.1. Strengths in the application of concept mapping

This review suggests the use of concept mapping in the measurement development and evaluation process has several notable strengths. First, concept mapping offers a *solid method for establishing content validity*. A clear conceptual framework is essential to the development of a valid scale or measure that has practical utility. Without a sound conceptual grounding it may be unclear if relevant elements have been identified, and thus reflect the phenomenon under study (Clark & Watson, 1995). Across the studies in this review, many noted the capacity of concept mapping to address such a challenge and yield an expanded set of representative items organized as a multidimensional network. It appears from this review that concept mapping is particularly well-suited to explicate a pattern of expected similarity among concepts within a broader content domain. The method yields much more than a list of items for populating a scale or measure, rather concept mapping produces a detailed relevant and representative conceptual structure based on a gradient of relational similarity (Trochim, 1989).

Second, concept mapping used in scale and measure development and evaluation *facilitates researcher decision-making*. In any research process on complex concepts there exist a tension between methodologically robust measurement and the extent to which concepts and their significance are understood and applied. Indeed, the more investigators know about the phenomenon in which they are interested and the implicit relationships that exist among the theoretical constructs, the better equipped they are to frame choices that yield develop reliable, valid, and practical measures (DeVellis, 2011). For many in this review, the presence of a clear conceptual framework that articulated the relationships between multiple concepts within the domain helped guide researcher decision-making, particularly in the selection of the content and hypothesized construct definitions. The systematic generating and structuring of the content domain enabled several

researchers to carry out item reduction and refinement processes using a large item pool to protect against compromises to construct validity (cf. Clark & Watson, 1995; Messick, 1993). Nearly all studies in this review promoted the strength of concept mapping in the process of identification, binning, and winnowing of items in the initial pool, although the specifics of how were not consistently provided.

Third, concept mapping used in measurement development and evaluation provides *insight into target population perspectives that are integrated a priori*. Several researchers in this review were explicit regarding the participatory significance of concept mapping to the measurement development process. For many, understanding how participants themselves conceptualize the relationship between attributes of specific items and higher-level constructs in advance of psychometric testing was important to interpreting the results of the statistical analyses. Despite the limited information provided by researchers regarding the rationale for concept mapping, given the method's participatory emphasis, we inferred the use of concept mapping was rooted in the intent to include multiple perspectives in the generation, structuring, and confirmation of the scale or measure content. Several researchers found concept mapping provided a structured, systematic means for meeting expectations regarding inclusive processes, such as those associate with the development of patient-informed measures. Clearly knowing in advance the preferences, values and judgments of target individuals increases the relevance and meaning of the instrument to the situation, greatly enhancing the accuracy and precision of the measurement.

Finally, concept mapping provides a *foundation for analytical and interpretative choices*. The studies reviewed here employed a variety of advanced psychometric techniques to examine the validity and reliability of the new instruments. Researchers avoided a simple empirical scale construction strategy and sought to maximize the utility of the psychometric analytic methods. To that end, concept mapping can establish strong and clear links between items within a broader theoretical domain. In situations where the conceptual underpinnings of target constructs are poorly understood, the application of a pre-existing models nor factor analysis of a list of attributes may be appropriate (Nunnally, 1978). Rarely do researchers conduct a measurement development study without an a priori hypothesis or theoretical grounding (Costello & Osborne, 2005). In each of the studies here, researchers began the measurement evaluation process with an explicated theory from which a set of core items can be systemically and empirically identified. Thus, sound judgments about the results of psychometric analyses can be enhanced by the a priori presence of a theoretical measurement pattern via concept mapping.

4.2. Limitations in the application of concept mapping and future directions

Notwithstanding the strengths of concept mapping in the measurement development process, there are several limitations in the application of the method based on this review. First, there exists a lack of consistent reporting of key elements in the process. A full description of the idea generation process, item reduction, and item construction varied across studies. Moreover, researchers tended to provide much less methodological detail for concept mapping that what was provided for the testing procedures. More consistent reporting of standardized features of concept mapping, such as those reported in Rosas and Kane (2012) are necessary to facilitate the review by others as to the quality of the application of the method. For example, although reporting of the stress value is inconsistent in the general concept mapping literature, its absence in the context of concept mapping informed scale development studies is particularly striking. Given the emphasis on concept mapping as tool for enhancing content validity, the lack stress value reporting as a standardized means for assessing the fit of the initial conceptual model upon which the instrument content and structure was based is noticeable. In addition, greater detail into the mechanics of how critical decisions using concept mapping are made can provide insights to future efforts to use the method in measurement development and evaluation. For instance, Windsor (2013) noted that concept mapping lacks a good statistical method for reducing the number of statements derived during the idea generation process. The lack of widely accepted procedures and methods for statement reduction have implications for pre- and post-map syntheses of content, and a more thorough approach to managing the content is warranted for sound measurement development. Providing detailed information regarding the origin of measures is a necessary prerequisite for new measures and a clear link between items and their theoretical domain needs to be established (Hinkin, 1995). Furthermore, the lack of documentation on approaches for item refinement and construction when moving from concept map to initial measurement tool development, limits our understanding of this critical decision-making step.

Second, it is also not clear based on this review, the degree to which the concept mapping process steps required tailoring or adjustment to maximize value to the scale development process. Greater specification as to which of the concept mapping steps, if any, that require adjustment is needed for researchers to understand how the approach can be applied without compromising key principles and practices. For example, given the burden to participants during some of the steps, the question of whether specific concept mapping tasks (e.g., sorting and rating) can be altered without affecting the measurement development process is unanswered. As it stands, the replication of any single measurement development study can be challenging, given the variety of methodological decisions needed for integrating concept mapping with development and testing procedures and the lack of formal guidance in the literature. More research and documentation that would lead to replicable processes and clearer guidance are needed.

Third, arguably the most significant limitation was the absence of the exploration of the linkage between the empirical results of the instrument (observed measurement pattern) and concept map (theoretical measurement pattern). The explanation of the degree to which the final psychometrically derived model comports with the original conceptual model (i.e., concept map) on which the measurement tool is based was limited. Given that the new instruments included in this review were based on the concept mapping process and results, this connection would seem to be significant and a discussion warranted. The correspondence between the theoretical constructs of interest and the methods of measurement that are used to operationalize them is not simply a novel idea – it is fundamental to psychometrically sound, accurate and relevant measurement. The importance of congruence between the measure and the measured cannot be overstated (McGrath, 2005).

Fourth, the studies were limited in detailing the rationale for the psychometric approach used to test and evaluate the quality of the instrument and the relationship to concept mapping as the core conceptualization technique. Regardless of the measurement models used, sound conceptualization procedures like concept mapping can be useful in maximizing the capacity of statistical procedures to yield useful estimates for interpreting psychometric properties. Even the most sophisticated analytical procedures cannot account for poor or inadequate a priori conceptualization of the content domain. Moreover, it is not clear how, if at all, the multi-dimensional nature of the concept mapping results informs the choice of analytical methods used to examine the psychometric

characteristics of the instrument, particularly those methods based on a unidimensional measurement framework (i.e., Rasch models). Beyond the generation of a large set of items, it is not apparent how the structural (e.g., clusters) and relational (e.g., dimensionality; proximity) elements of concept shape analytical decisions. Thus, in the context of the analytical steps toward content and construct validation, concept mapping appears to be underexplored and future work should strive to address these broader epistemological questions.

Finally, the level of adoption of the new tools produced in these studies is unknown. However, future analysis might include citation metrics as possible rough estimates of adoption. For the future, a more detailed description of the overall quality and utilization of measures born out of concept mapping should be documented.

4.3. Summary

As an integral part of traditional measurement development and psychometric testing processes, the overall value of concept mapping can be assessed in the context of three general analytical steps toward content and construct validation (Pedhazur & Schmelkin, 2013; Simms, 2008). First, logical analysis involves the consideration of the content and logical structure of the domain construct and involves critical thinking and judgment. A structured consultative approach employing concept mapping can illuminate not only the content from which a scale or measure might be constructed, but the structure of the interrelationships among elements. Such knowledge is important for making decisions about what is and is not core to the measurement pattern. Comprehensiveness of the content and clarity of the conceptual origin is critical to defining components of any new measurement tool and is often found to be an issue (Hinkin, 1995; Simms, 2008). Second, internal structural analysis involves the generation of evidence that a set of items co-vary based on a priori structural hypotheses to determine relevant versus irrelevant variance. Following the specification of an initial model, and subsequently the re-specification and confirmation of the model through psychometric analysis, this iterative analytical process relies on a solid understanding of structural relations among items parallel to other structural relations in the domain of interest (Loevinger, 1957). The use of concept mapping in the explication of a theoretical measurement pattern enables researchers to employ a strategy to examine the structural fidelity. Finally, cross structural-analysis requires the generation of evidence supporting hypothesized relationships between the construct and other constructs. In this step, the development of a priori relational hypotheses as a basis for ongoing examination of observed results and operating theory of the construct(s) is emphasized. While it may seem like concept mapping may be limited in this analytical step, as a rigorous structured consultation and conceptualization process, concept mapping can suggest a priori hypotheses and provide a firm foundation for building the evidence of validity for a wide range of potential interpretations and applications (Buchbinder et al., 2011).

There are practical, methodological, and epistemological consequences of poor measurement, including deficient/contaminated measure, measurement model mis-specification, and weak theoretical rationale for hypotheses (MacKenzie, 2003). Developing sound instruments is a difficult, time-consuming, and costly enterprise, especially considering the complexity and challenge of establishing construct validity of a new measurement tool (Schmitt & Klimoski, 1991). While what we have reported on here may seem to be typical considerations for the measurement development process, we believe concept mapping as an integrated method drives these considerations to the forefront. Dialogue with targets, robust psychometric testing, and assessment of the final content with the original conceptual framework is a continuous process that frames the generation of evidence related to the validity and utility of the instrument (Buchbinder et al., 2011). As has been illuminated in this review, concept mapping can occupy a key methodological role in ongoing processes to evaluate the measurement characteristics and quality of newly developed instruments.

References[1]

Armstrong, N. P., & Steffen, J. J. (2009). The recovery promotion fidelity scale: Assessing the organizational promotion of recovery. *Community Mental Health Journal*, *45*(3), 163–170.

Battterham*, R., Southern, D., Appleby, N., Elsworth, G., Fabris, S., Dunt, D., & Young, D. (2002). Construction of a GP integration model. *Social Science & Medicine*, *54*, 1225–1241.

Behar*, L. B., & Hydaker, W. M. (2009). Defining community readiness for the implementation of a system of care. *Administration and Policy in Mental Health and Mental Health Services Research*, *36*(6), 381–392.

Behfar*, K. J., Mannix, E. A., Peterson, R. S., & Trochim, W. M. (2011). Conflict in small groups: The meaning and consequences of process conflict. *Small Group Research*, *42*(2), 127–176.

Beijersbergen*, M. D., Asmoredjo, J. K., Christians, M. G., & Wolf, J. R. (2014). Psychometric properties of the consumer quality index to assess shelter and community care services. *The European Journal of Public Health*1–7. http://dx.doi.org/10.1093/eurpub/cku195.

Buchbinder, R., Batterham, R., Elsworth, G., Dionne, C. E., Irvin, E., & Osborne, R. H. (2011). A validity-driven approach to the understanding of the personal and societal burden of low back pain: Development of a conceptual and measurement model. *Arthritis Research &Therapy*, *13*(5), R152. http://dx.doi.org/10.1186/ar3468.

Butler*, S. F., Budman, S. H., Fernandez, K., & Jamison, R. N. (2004). Validation of a screener and opioid assessment measure for patients with chronic pain. *Pain*, *112*(1), 65–75.

Butler*, S. F., Budman, S. H., Fernandez, K. C., Houle, B., Benoit, C., Katz, N., & Jamison, R. N. (2007). Development and validation of the current opioid misuse measure. *Pain*, *130*(1), 144–156.

Campbell, D. T. (1966). Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik* (pp. 81–106).Oxford: Holt, Rinehart and Winston.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. *New Directions for Program Evaluation*, *31*, 67–77.

Carpenter*, B. D., Van Haitsma, K., Ruckdeschel, K., & Lawton, M. P. (2000). The psychosocial references of older adults: A pilot examination of content and structure. *The Gerontologist*, *40*(3), 335–348.

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research*, *23*(2), 485–493.

Ciciriello*, S., Buchbinder, R., Osborne, R. H., & Wicks, I. P. (2014). Improving treatment with methotrexate in rheumatoid arthritis: Development of a multimedia patient education program and the MiRAK, a new instrument to evaluate methotrexate-related knowledge. *Seminars in Arthritis and Rheumatism*, *43*(4), 437–446.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in scale development. *Psychological Assessment*, *7*(3), 309–319.

Converse, J. M., & Presser, S. (1986). 63 *Survey questions: Handcrafting the standardized questionnaire*, *vol. 7*, Thousand Oaks, CA: Sage.

Corcoran*, K. (2005). The Oregon mental health referral checklists: Concept mapping the mental health needs of youth in the juvenile justice system. *Brief Treatment and Crisis Intervention*, *5*(1), 9–18.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7) [Retrieved from] http://pareonline.net/getvn.asp?v=10&n=7.

Conrad*, K. J., Iris, M., Ridings, J. W., Langley, K., & Wilber, K. H. (2010). Self-report measure of financial exploitation of older adults. *The Gerontologist*, *50*(6), 758–773.

Conrad*, K. J., Iris, M., Ridings, J. W., Langley, K., & Anetzberger, G. J. (2011). Self-report measure of psychological abuse of older adults. *The Gerontologist*, *51*(3), 354–366.

Conrad*, K. J., Iris, M., Ridings, J. W., Rosen, A., Fairman, K. P., & Anetzberger, G. J. (2011). Conceptual model and map of psychological abuse of older adults. *Journal of Elder Abuse & Neglect*, *23*(2), 147–168.

Conrad*, K., Ridings, J., Iris, M., Fairman, K., Anetzberger, G., & Rosen, A. (2011). Conceptual model and map of financial exploitation of older adults. *Journal of Elder Abuse and Neglect*, *23*(3) .

[1] Astricks denotes studies included in the review.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Davison, M. L. (1983). *Multidimensional scaling*. New York: John Wiley and Sons.

DeVellis, R. F. (2011). *Scale development: Theory and applications*, *Vol. 26*, Thousand Oaks, CA: Sage Publications.

de Kok*, M., Scholte, R. W., Sixma, H. J., van der Weijden, T., Spijkers, K. F., van de Velde, C. J., . . . von Meyenfeldt, M. F. (2007). The patient's perspective of the quality of breast cancer care: The development of an instrument to measure quality of care through focus groups and concept mapping with breast cancer patients. *European Journal of Cancer*, *43*(8), 1257–1264.

de Kok*, M., Sixma, H. J. M., van der Weijden, T., Kessels, A. G. H., Dirksen, C. D., Spijkers, K. F. J., . . . von Meyenfeldt, M. F. (2010). A patient-centered instrument for assessment of quality of breast cancer care: Results of a pilot questionnaire. *Quality and Safety in Health Care*, *19*(6), 1–8. http://dx.doi.org/10.1136/qshc.2007.025890.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah NJ: Lawrence Erlbaum Associates.

Everitt, B. (1980). *Cluster analysis*, 2nd ed. New York: Halstead Press.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286–299.

Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology*. Thousand Oaks, CA: Sage Publications.

Galvin, P. F. (1989). Concept mapping for planning and evaluation of a big brother/big sister program. *Evaluation and Program Planning*, *12*(1), 53–57.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, *21*(5), 967–988.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*(1), 104–121.

Iris*, M., Ridings, J., & Conrad, K. (2010). The development of a conceptual framework for understanding elder self-neglect. *The Gerontologist*, *50*(3), 303–315.

Iris, M., DeBacker, N. A., Benner, R., Hammerman, J., & Ridings, J. (2012). Creating a quality of life assessment measure for residents in long term care. *Journal of the American Medical Directors Association*, *13*(5), 438–447.

Iris*, M., Conrad, K., & Ridings, J. (2014). Observational measure of elder self-neglect. *Journal of Elder Abuse & Neglect*, *26*(4), 365–397.

Jordan*, J. E., Buchbinder, R., Briggs, A. M., Elsworth, G. R., Busija, L., Batterham, R., & Osborne, R. H. (2013). The Health Literacy Management Scale (HeLMS): A measure of an individual's capacity to seek, understand and use health information within the healthcare setting. *Patient Education and Counseling*, *91*(2), 228–235.

Kane, M., & Trochim, W. M. K. (2007). *Concept mapping for planning and evaluation*. Thousand Oaks, CA: Sage Publications.

Kane, M., & Trochim, W. M. K. (2009). Concept mapping for applied social research. In L. Bickman, & D. Rog (Eds.), *The sage handbook of applied social research methods* (pp. 435–474).Thousand Oaks, CA: Sage Publications.

Kline, R. B. (2006). *Principles and practice of structural equation modeling*, 2nd ed. New York: Guilford.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Loevinger, J. (1957). Objective tests as instruments of psychological theory: Monograph Supplement 9. *Psychological Reports*, *3*(3), 635–694.

Luke*, D. A., Calhoun, A., Robichaux, C. B., Elliott, M. B., & Moreland-Russell, S. (2014). The program sustainability assessment tool: A new instrument for public health programs. *Preventing Chronic Disease*130184. http://dx.doi.org/10.5888/pcd11.130184.

MacKenzie, S. B. (2003). The dangers of poor construct conceptualization. *Journal of the Academy of Marketing Science*, *31*(3), 323–326.

McGrath, R. E. (2005). Conceptual complexity and construct validity. *Journal of Personality Assessment*, *85*(2), 112–124.

Messick, S. (1993). Validity, In R. L. Linn (Ed.), *Educational measurement* (pp. 105–146).2nd ed. Phoenix, AZ: American Council on Education and the Oryx Press.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (Eds.). (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage Publications.

Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill Book Company.

Osborn, A. F. (1957). *Applied imagination*. New York: Scribner.

Osborne*, R. H., Elsworth, G. R., & Whitfield, K. (2007). The Health Education Impact Questionnaire (heiQ): An outcomes and evaluation measure for patient education and self-management interventions for people with chronic conditions. *Patient Education and Counseling*, *66*(2), 192–201.

Osborne*, R. H., Norquist, J. M., Elsworth, G. R., Busija, L., Mehta, V., Herring, T., & Gupta, S. B. (2011). Development and validation of the influenza intensity and impact questionnaire (FluiiQ™). *Value in Health*, *14*(5), 687–699.

Osborne*, R. H., Batterham, R. W., Elsworth, G. R., Hawkins, M., & Buchbinder, R. (2013). The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). *BMC Public Health*, *13*, 658. http://dx.doi.org/10.1186/1471-2458-13-658.

Pedhazur, E. J., & Schmelkin, L. P. (2013). *Measurement, design, and analysis: An integrated approach*. New York: NY. Psychology Press.

Rosas*, S. R., & Camphausen, L. C. (2007). The use of concept mapping for scale development and validation in evaluation. *Evaluation and Program Planning*, *30*(2), 125–135.

Rosas, S. R., & Kane, M. (2012). Quality and rigor of the concept mapping methodology: a pooled study analysis. *Evaluation and Program Planning*, *35*(2), 236–245.

Rosas*, S. R., Behar, L. B., & Hydaker, W. M. (2014). Community Readiness within Systems of Care: the Validity and Reliability of the System of Care Readiness and Implementation Measurement Scale (SOC-RIMS), The Journal of Behavioral Health Services &. *Research*1–20.

Schell*, S. F., Luke, D. A., Schooley, M. W., Elliott, M. B., Herbers, S. H., Mueller, N. B., & Bunger, A. C. (2013). Public health program capacity for sustainability: A new framework. *Implementation Science*, *8*(1) . http://dx.doi.org/10.1186/1748-5908-8-15.

Schmitt, N. W., & Klimoski, R. J. (1991). *Research methods in human resources management*. Cincinnati: South-Western Publishing.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, *19*, 385–417.

Sepúlveda Carrillo, G. J., Meneses Báez, A. L., & Goldenberg, P. (2014). Content validity: The human papillomavirus vulnerability questionnaire. *Enfermería Global*, *13*(3), 211–239.

Shorkey, C., Windsor, L. C., & Spence, R. (2008). Assessing culturally competent chemical dependence treatment services for Mexican Americans. *The Journal of Behavioral Health Services & Research*, *36*(1), 61–74.

Shorkey, C., Windsor, L. C., & Spence, R. (2009). Systematic assessment of culturally competent chemical dependence treatment services for African Americans. *Journal of Ethnicity in Substance Abuse*, *8*(2), 113–128.

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass*, *2*(1), 414–433.

Southern*, D. M., Young, D., Dunt, D., Appleby, N. J., & Batterham, R. W. (2002). Integration of primary health care services: Perceptions of Australian general practitioners, non- general practitioner health service providers and consumers at the general practice–primary care interface. *Evaluation and Program Planning*, *25*(1), 47–59.

Streiner, D. L., Norman, G. R., & Cairney, J. (2014). *Health measurement scales: A practical guide to their development and use*, 5th ed. Oxford, UK: Oxford University Press.

Tabachnick, B. G., & Fidell, L. S. (2001). *Multivariate statistics*. Needham Heights, MA: Allyn & Bacon.

Trochim, W., & Linton, R. (1986). Conceptualization for evaluation and planning. *Evaluation and Program Planning*, *9*, 289–308.

Trochim, W. M. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review*, *9*(5), 575–604.

Trochim, W. M. K. (1989). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning*, *12*(1), 1–16.

Trochim, W. M. K. (1993). The reliability of concept mapping. *Paper presented at the annual conference of the American Evaluation Association* November.

Van Haitsma*, K., Curyto, K., Specor, A., Towsley, G., Kleban, M., Carpenter, B., . . . Koren, M. J. (2012). The preferences for everyday living inventory: Scale development and description of psychosocial preferences responses in community-dwelling elders. *The Gerontologist*, *53*(4), 582–595.

Velozo, C. A., Seel, R. T., Magasi, S., Heinemann, A. W., & Romero, S. (2012). Improving measurement methods in rehabilitation: Core concepts and recommendations for scale development. *Archives of Physical Medicine and Rehabilitation*, *93*(8), S154–S163.

Wallace*, L. S., Wexler, R. K., Miser, W. F., McDougle, L., & Haddox, J. D. (2013). Development and validation of the patient opioid education measure. *Journal of Pain Research*, *6*, 663–681.

White, K. S., & Farrell, A. D. (2001). Structure of anxiety symptoms in urban children: Competing factor models of Revised Children's Manifest Anxiety Scale. *Journal of Consulting and Clinical Psychology*, *69*(2), 333.

Windsor, L. C. (2013). Using concept mapping in community-based participatory research: A mixed methods approach. *Journal of Mixed Methods Research*, *7*(3), 274–293.

Wolfinbarger*, M., & Gilly, M. C. (2003). eTailQ: Dimensionalizing, measuring and predicting etail quality. *Journal of Retailing*, *79*(3), 183–198.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research a content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838.

van Nieuwenhuizen*, C., Schene, A. H., Koeter, M. W. J., & Huxley, P. J. (2001). The lancashire quality of life profile: Modification and psychometric evaluation. *Social Psychiatry and Psychiatric Epidemiology*, *36*(1), 36–44.

van der Eijk*, I., Sixma, H., Smeets, T., Veloso, F. T., Odes, S., Montague, S., . . . Russel, M. (2001). Quality of health care in inflammatory bowel disease: Development of a reliable questionnaire (QUOTE-IBD) and first results. *The American Journal of Gastroenterology*, *96*(12), 3329–3336.

vander Waal, M. A. E., Casparie, A. F., & Lako, C. J. (1996). Quality of care: A comparison of preferences between medical specialists and patients with chronic diseases. *Social Science & Medicine*, *42*(5), 643–649.