THE PSYCHOMETRIC PROPERTIES OF INSTRUMENTS USED TO ASSESS ANXIETY IN OLDER ADULTS

Zoé Therrien-Poirier

Dissertation submitted to the

Faculty of Graduate and Postdoctoral Studies

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Clinical Psychology

University of Ottawa

Ottawa, Ontario

2012

Abstract

With the growing number of older adults in the general population, there is also a concomitant rise in the number of older adults who require mental health services, making the measurement of psychological conditions in later life a priority. However, due to a lack of measures created for older adults, researchers and clinicians must often rely on measures created for younger populations. Three studies were designed to add to the field of evidence-based assessment and determine which anxiety measures possess strong evidence when used with older adults to warrant their use with this specific population. In the first study, I systematically reviewed the literature to identify the anxiety measures most commonly used with older adults. I reviewed each measure to examine its psychometric properties (e.g., internal consistency, test-retest reliability, inter-rater reliability, concurrent and discriminant validity) and the availability of age-appropriate norms in order to evaluate whether the instruments are appropriate for use with older adults. In the second study, I conducted a reliability generalization meta-analysis to estimate the mean reliability of each commonly used anxiety measure identified in the first study. Finally, in the third study, I examined whether the anxiety measures commonly used with an older population can be consistently and accurately categorized as evidence-based. The literature review and the reliability generalization study both revealed that most of the most commonly used measures lacked sufficient evidence to warrant their use with older adults. However, three measures (Beck Anxiety Inventory, Penn State Worry Questionnaire, and Geriatric Mental Status Examination) showed psychometric properties sufficient to justify the use of these instruments when assessing anxiety in older adults. In addition, two measures developed specifically for older adults (Worry Scale and Geriatric Anxiety Inventory) were also found to be appropriate for use with older adults. This suggests that based on their overall level of reliability and previous

psychometric evidence, both researchers and clinicians assessing anxiety in a geriatric population should consider these measures as likely to be the best currently available.

Statement of Co-Authorship

The three manuscripts included in this dissertation were prepared in collaboration with my dissertation supervisor. I was primary author and Dr. John Hunsley was the secondary author for the first manuscript, entitled “Assessment of Anxiety in Older Adults: A Systematic Review of Commonly Used Measures”, the second manuscript entitled “Assessment of Anxiety in Older Adults : A Reliability Generalization Meta-Analysis of Commonly Used Measures”, and the third manuscript entitled “Comparing Approaches for Categorizing Measure Reliability in the Assessment of Anxiety in Older Adults”. As the primary author on all manuscripts, I was responsible for the conceptualization of the research question and methods, planning and execution of statistical analyses, and preparation of manuscripts. Dr. Hunsley provided guidance and assistance in all aspects of the project, especially in the refinement of the research methods, and editing of the manuscripts.

Publication Statement

Study 1, Assessment of Anxiety in Older Adults: A Systematic Review of Commonly Used Measures, was published in Aging and Mental Health: Therrien, Z., & Hunsley, J. (2012). Assessment of anxiety in older adults: A systematic review of commonly used measures. *Aging and Mental Health, 16,* 1-16.

Study 2, Assessment of Anxiety in Older Adults: A Reliability Generalization Meta-Analysis of Commonly Used Measures, is currently under press in Clinical Gerontologist: Therrien, Z., & Hunsley, J. (In press). Assessment of Anxiety in Older Adults: A Reliability Generalization Meta-Analysis of Commonly Used Measures. *Clinical Gerontologist*.

Study 3, Comparing Approaches for Categorizing Measure Reliability in the Assessment of Anxiety in Older Adults, is currently under review in Administration and Policy in Mental Health and Mental Health Services Research.

As requested at the dissertation defence, some modifications have been made to the final version of the dissertation. Smaller changes were directly in the text of the dissertation and therefore some minor differences will be found between this dissertation and the first two studies that were published in *Aging and Mental Health* and in *Clinical Gerontologist*. For more substantive changes, footnotes were added to the text in order to maintain the integrity of the two published studies.

Acknowledgements

Foremost, I would like to thank my supervisor, Dr. John Hunsley, for the continuous support of my Ph.D. study and research, for his patience, motivation, enthusiasm and immense knowledge. Your guidance helped me in all the time of research and writing of this thesis and will continue to guide me in my future endeavors. Thank you for teaching me what I should aim for as a researcher. Beside my supervisor, I would like to thank the rest of my thesis committee, Drs Jean Grenier, John Lyons and Vanessa Taler, for their encouragement, insightful comments and hard questions. I must also acknowledge Dr. Dwayne Schindler who listened to my many questions and provided me with statistical advice at times of critical need.

I am forever thankful to my undergraduate professors at the Université de Hearst. Their enthusiasm and encouragement were the reason why I decided to pursue graduate studies. I doubt I will ever be able to convey my appreciation fully. A special thank you to Dr. Jean-Pierre Bergevin who providing advice and assistance many times over the last years. You were and remain my best role model for a psychologist, mentor and teacher. Thank you to Professor Pierre Bouchard for giving me my first experiences as a researcher. I will be forever grateful for your assistance with writing letters, giving wise advice and helping me with various applications.

I would like to thank my dear friends and colleagues at the University of Ottawa, especially Isabelle Arès, Monic Gallien and Mylène Laforest for the stimulating discussions, the sleepless nights we were working together before deadlines and for all the laughs we shared in the last five years. I am forever grateful to have shared this experience with you and look forward to the many years of friendship ahead of us. A special thank you to my colleague and friend Éric Thériault. Your constant support and encouragement has pushed me through the most difficult days. I also thank my friends outside of the program for helping me get through difficult times and for all the emotional support, camaraderie, entertainment and care they provided.

I would also like to express eternal appreciation towards my family and family-in-law for always being there for me no matter where I am. A special thank you to my mom for her unceasing encouragement, unconditional support and patience. Finally, I thank my partner Jonathan for blindly believing in me. You have done more sacrifices in the last years than I may ever be able to compensate for the many years we have ahead of us. This would not of been possible if not for you.

Table of Contents

List of Tables

The Psychometric Properties of Instruments Used to Assess Anxiety in Older Adults

Zoé Therrien

University of Ottawa

The Reliability of Instruments Used to Assess Anxiety in Older Adults

Over the last decade, there has been a growing emphasis on the importance of evidence-based mental health practice (American Psychological Association Presidential Task Force on Evidence Based Practice, 2006; Ayers, Sorrell, Thorp, & Wetherell, 2007; Institute of Medicine, 2001). According to the Institute of Medicine (2001), evidence-based practice in health care services integrates information derived from the best research evidence, clinical expertise and the patient's values when considering options for health care services. This framework is based on the acknowledgement that positive outcomes are less likely if the chosen health care service does not have a research base that shows the potential to improve the client's functioning (Forman, Fagley, Steiner, & Schneider, 2009). The evidence-based practice movement has been endorsed through reports released by several scientific organizations in the field of health and mental health (e.g., APA, Division 12, 2004; Institute of Medicine, 2001). A growing body of research has also concentrated on the identification of efficacious psychological interventions based on research (e.g., Ayers et al., 2007) and the establishment of guidelines that would help in determining whether an intervention can be considered evidence-based (e.g., Chambless & Hollon, 1998).

The effort to identify, describe and disseminate information about interventions supported by strong scientific research is undoubtedly critical for the advancement of professional psychology. However, given the longstanding importance of accurate measurement in the field of psychology, it seems incongruous that evidence-based initiatives began with treatments rather than with assessment (Hunsley & Mash, 2007). Indeed, without accurate assessment data, clinicians cannot clearly evaluate a patient's level of functioning; without this information, the development of solid case formulations (including diagnostic considerations) is not possible, and without high quality case formulations, making informed treatment choices is nearly impossible.

Moreover, in research contexts, assessment measures are used to select participants and evaluate treatment; thus, they are critical for developing evidence-based treatments (Cohen et al., 2008). In sum, as the identification and evaluation of evidence-based treatments rests entirely on assessment data, overlooking the quality of assessment instruments places the movement for evidence-based psychological practice in jeopardy.

Fortunately, over the past decade, a modest but noticeable shift has occurred in the assessment literature, with increased attention being paid to the way in which psychological assessment is conducted and used to guide practice. This led to the development of evidence-based assessment (EBA), an approach to clinical evaluation that emphasizes the use of research and theory to guide the selection of constructs, the methods and measures used in the assessment as well as the whole process of assessment (Hunsley & Mash, 2007). In doing so, clinicians need to recognize the importance of assessment measures presenting with strong psychometric properties and remember that the assessment process is a decision-making task in which one needs to decide how to best gather and integrate data in order to formulate and test hypotheses (Mash & Hunsley, 2005).

Although research on evidence-based assessment is increasing, research on the assessment of older adults is much less developed (Ayers et al., 2007). With the growing number of older adults in the general population, there is also a concomitant rise in the number of older adults who require mental health services, making the measurement of psychological conditions in later life a priority. However, the psychological assessment of older adults is complicated by many factors, such as the common comorbidity of mental and physical health problems, the presence of various medication, as well as age-related changes (Edelstein et al., 2008). These differences make it unlikely that a measure of psychological functioning will perform accurately across the lifespan. But, due to a lack of measures created for older adults, researchers and clinicians must

often rely on measures created for younger populations. It is therefore increasingly important that measures used with older adults be tested for reliability and validity with older adults and to establish age appropriate norms (Dennis, Boddington & Funnell, 2007).

An Overview of the Dissertation

With this background in mind, the primary goal of this dissertation is to add to the field of evidence-based assessment and determine which anxiety measures possess strong psychometric evidence when used with older adults to warrant their use with this specific population. In the first study, literature searches were conducted to identify the anxiety measures that are commonly used with older adults. Each of these measures was reviewed in order to examine the psychometric properties (e.g., internal consistency, test-retest reliability, inter-rater reliability, concurrent and discriminant validity) and the availability of age-appropriate norms in order to evaluate whether the instruments are appropriate for use with older adults. Based on the same set of measures, the second study adds to the systematic review conducted in the first study by using a meta-analytic method known as reliability generalization (RG) to estimate the mean reliability of each measure. Finally, with the availability of a thorough evaluation of the psychometric soundness of each reviewed anxiety measure, the third study examines whether the anxiety measure commonly used with an older population can be consistently and accurately categorized as evidence-based. This will be done by assessing different existing categorization systems used to designate whether a measure can be considered as evidence-based and comparing the results of this classification process to the evidence available from the meta-analysis conducted in the second study.

Definitions of Anxiety and Related Terms

As many terms are used in the literature to describe the experience of anxiety, there seems to be some considerable confusion and inaccuracy surrounding the common use of the term. It is

therefore important to define and clearly distinguish terms such as *anxiety, fear and worry* in order to offer guidance for research and treatment of anxiety. Beck, Emery, and Greenberg (1985) emphasized the cognitive nature of fear and defined it as the appraisal of actual or potential danger in a specific situation and described anxiety as the affective response that is generated by the fear. On the other hand, in his influential volume on anxiety disorders, Barlow (2002) focused on the neurobiological and behavioral features of the constructs and described anxiety as: "a future-oriented emotion, characterized by perceptions of uncontrollability and unpredictability over potentially aversive events and a rapid shift in attention to the focus of potentially dangerous events or one's own affective response to these events" (p.104). He described fear as a "primitive alarm in response to present danger, characterized by strong arousal and action tendencies" (p.104).

Based on these considerations, I will use the definitions suggested by Clark and Beck (2010) that encompasses both the cognitive and neurobiological features of the constructs. Fear can be defined as the "primitive automatic neurophysiological state of alarm involving the cognitive appraisal of imminent threat or danger to the safety and security of an individual" (p.5). Anxiety can be defined as a "complex cognitive, affective, physiological, behavioral response system that is activated when anticipated event or circumstances are deemed to be highly aversive because they are perceived to be unpredictable, uncontrollable events that could potentially threaten the vital interests of an individual" (p. 5). Therefore fear is a core process common to all anxiety disorders as it is the basic automatic appraisal of threath whereas anxiety is a more persistent state of threat that, in addition to fear, includes other cognitive factors such as perceived uncontrollability and uncertainty.

Borkovec, Robinson, Pruzinsky and DePree (1983) were one of the first to distinguish worry from other constructs. They offered a definition that has become widely accepted in the

research literature: “Worry is a chain of thoughts and images, negatively affect-laden and relatively uncontrollable.” (p.10). They specify that worrying represents an attempt to mentally problem solve an issue when there is a possibility of one or more negative outcome. Although this definition persists, a more complicated picture has emerged and there appears to be a consensus that worry serves as an avoidant coping function by suppressing somatic and negative emotional responses to internally represented threat cues that is negatively enforced by reductions in emotional reactivity (Sibrava & Borkovec, 2006). However, because individuals do not process their distress other than in abstract terms, they experience ongoing distress and continue to use worry to reduce this distress (Dugas & Robichaud, 2007).

Over the last several decades, clinical research on anxiety has recognized that there are a number of specific subtypes to clinically significant anxiety that cluster under the category of anxiety disorders. These include generalized anxiety disorder, phobic disorders, obsessive-compulsive disorder and post-traumatic stress disorder. It is important to note that epidemiological studies have often found a high rate of comorbidity between anxiety disorders and depression, an affective or mood disturbance that is accompanied by both psychological (i.e. sadness, generalized withdrawal of interest, diminished self-esteem) and physical symptoms (i.e. fatigue, loss of energy). This is not surprising as research indicates that there is considerable overlap between anxiety and depressive symptoms, as measured by commonly used self-report questionnaires (e.g. Seligman & Ollendick, 1998). For example, symptoms such as sleep disturbance, appetite changes, nonspecific cardio-pulmonary or gastrointestinal complaints, difficulty concentrating, irritability and fatigue or lack of energy are shared by both anxiety and depression. However, there are also a number of symptoms that help differentiate the two conditions. Patients often state that anxiety symptoms occur first followed by despair over their inability to cope. Anxiety itself is characterized by nervousness, worry and fear, and with

associated insomnia in the evening. It is usually worse in the evening as life stressors progress (Lieerman & Stein, 2009). Depression, on the other hand, is a low-energy state, with loss of drive and motivation. It is often associated with early morning awakening and is frequently worse in the morning.

Over the past two decades, structural models of anxiety and depression have emerged to explain the shared and unique features of these highly related and commonly comorbid conditions. Of these models, Clark and Watson's (1991) tripartite model of anxiety and depression has received the greatest amount of empirical attention. The tripartite model posits a general distress or negative affect (NA) factor that is shared by both anxiety and depression that accounts for symptom overlap and comorbidity. NA represents the extent to which an individual feels upset or unpleasantly engaged, rather than peaceful. It also indicates a sense of high objective distress and encompasses a variety of affective states including feeling angry, guilty, afraid, sad and worried. The model also suggests that although depression and anxiety share a common component of NA, they can be differentiated by two constructs: positive affect (PA) and physiological hyperarousal (PH). Individuals with symptoms or diagnoses of depression tend to exhibit low levels of PA and high levels of NA, whereas individuals with anxiety disorders tend to exhibit high levels of PH as well as high levels of NA. PA represents pleasurable engagement with the environment and encompasses moods such as energetic, active, and interested while the absence of PA reflects terms such as tired and fatigue. PH includes somatic symptoms such as shortness of breath, dizziness and lightheadedness. Although some authors have refined the complexities of the anxious-arousal dimension (e.g., the integrated-hierarchical model of Mineka, Watson, & Clark, 1998), the basic structure of both specific and non-specific symptoms has a strong body of support.

Mental Health in Older Adults

Policy-makers and researchers in most developed countries have accepted the age of 65 years as the definition of older adult (Erber, 2010). The association of age 65 with the start of older adulthood can be traced to the Social Security system that was established by the U.S. government in 1935 in order to provide economic security in the form of a monthly pension to older adults when retired from the workforce. As workers became eligible for Social Security pension benefits once they reached the age of 65, it became and remains the marker of older adulthood. In many parts of the developing world, other socially constructed meaning of age are more significant, such as the roles assigned to older adults. In some cases it is the loss of roles accompanying physical decline that is significant in defining old age. In the remainder of the dissertation, the chronological age of 65 will be used to define the term *older adults*.

Contrary to popular beliefs, a high proportion of older adults report being happy and leading fulfilling lives and do not seem to experience as much distress as do younger adults (Charles, Mather & Cartensen, 2003; Hinrichsen & Dick-Siskin, 2000). Older people today are healthier, wealthier and better educated than in previous generations, and this trend will likely continue in future generations. Although older adults, especially women, are disproportionately likely to live below the poverty line, the economic status of older adults has improved in recent years (Erber, 2010). Many pursue new activities while enjoying close relationships with family members and friends, and most cope well with illnesses, physical limitations and the loss of loved ones (Erber, 2010). Research on psychological functioning suggest that, in the absence of dementia, declines in fluid intelligence, memory and other abilities are relatively small and do not interfere with the ability to function in daily life (Zarit & Zarit, 2007). Despite these small declines in older adulthood, gains may be used to offset them. For example, gains in crystallized abilities may compensate for losses in fluid abilities and similarly, gains in the pragmatics of intelligence may compensate for losses older adults may experience in the mechanics of

intelligence (Erber, 2010). As part of this optimistic picture of aging, it is not surprising that approximately one-third of Canadians aged 65 and over report high levels of life satisfaction (Statistics Canada, 2005).

For some people, however, the challenges that come with aging, such as physical ailments, mobility difficulties, chronic pain and cognitive impairments can be debilitating and have a significant impact on their daily functioning. Other challenges such as retirement, widowhood, and the loss of relationships through death can also affect the social and emotional functioning of older adults and lead to diagnosable psychiatric disorders that need mental health services (Erber, 2010). Several decades ago, the estimated prevalence rate of mental disorders in older adults ranged from 12 to 22% (Dye, 1978), but more recently, estimates appear to be slightly higher and range from 18% to 28% (American Association of Geriatric Psychiatry, 2008; Gatz, 1995; Gatz & Finkel, 1995; Qualls & Smyer, 2006). Gatz and Smyer (2001) estimated that, when including conditions such as emotional dysfunction and cognitive impairment, 22% of combined community-living and institutional-living adults aged 65 years and older suffer some type of mental disorder. There are indications that the number of older adults with mental health problems is likely to increase (Knight, Kaskie, Shurgot & Dave, 2006). By 2026, one Canadian in five will have reached age 65 which consequently leads into a greater number of older adults presenting with mental health difficulties. This is due, in part, to the "baby boom generation" experiencing high rates of depression, anxiety and substance abuse that might persist into old age (Knight, 2004; Parmalee, 2007).

The most common disorders in older adults, in order of prevalence, are (a) anxiety disorders (ranging from 1.2% to 28%; Bryant, Jackson & Ames, 2009) such as phobias and generalized anxiety disorder, (b) severe cognitive impairment (ranging from 6.4% to 16%; Canadian Study of Health and Aging Working Group, 1994; Lobo et al., 2000) including

Alzheimer's disease, and (c) mood disorders such as depression (ranging from 1.8 to 13.2%; Mitchell, Rao, & Vaze, 2010). Although these prevalence rate are similar to what is found in the general population, studies suggest that mental disorders in older adults are probably under-reported as the assessment can be influenced by important biological, psychological and social changes associated with aging. Many older adults present persistent symptoms of a specific mental health disorders without meeting the full criteria as described in the Diagnostic and Statistical Manual of Mental Disorders (Beck & Averill, 2004; Streiner, Cairney & Velahuizen, 2006; Turnbull, 1989). Although subsyndromal, the consequences of these conditions may be as significant as those experienced with diagnosable conditions, but clinically significant but not diagnosable conditions are often not included in estimated prevalence rates (D'Hudson & Saling, 2010). Additionally, the combination of medical and psychological problems is more common among older adults (Zarit & Zarit, 2007). A high rate of medical and psychological comorbidity and associated pharmacological treatment can complicate, mimic or mask psychopathology, making the diagnosis more complex. Finally, changes due to normal aging can also make the diagnosis and assessment of mental disorders in late life challenging, as a specific symptom such as sleeping difficulties could be caused either by normal aging, by a physical illness or by a mental illness (Lenze & Wetherell, 2009).

Anxiety in Older Adults

The empirical literature on anxiety prevalence suggests that it has become a widespread problem in later life that is more prevalent than depression. In a recent review of the literature, Bryant, Jackson, and Ames (2009) found 2-month prevalence estimates ranging from 1.2 to 15% in community samples of older adults and from 1 to 28% in clinical samples of older adults. Furthermore, among hospitalized geriatric patients, the prevalence in the preceding 3 days is estimated to be as high as 43% and is known to remain high after discharge (Kvall,

Macijauskiene, Engedal, & Laake, 2001). The prevalence of subsyndromal anxiety symptoms is even higher, with prevalence rates ranging from 15 to 52.3% in community samples and 15 to 56% in clinical samples of older adults (Bryant et al., 2009). Older adults who present symptoms of anxiety are as negatively affected in their quality of life as those meeting the diagnostic criteria for an anxiety disorder, and the impact of this sub-threshold anxiety can cause significant disability that requires mental health services (DeBeurs et al., 1999; D'Hudson & Sailing, 2010; Segal, June, Payne, Coolidge & Yochim, 2010). Overall, the presence of anxiety in older adults is associated with poorer quality of life, increased health services use, poorer health and self-perception of health, increased subjective distress, decreased life satisfaction and decreased physical activities (Segal et al., 2010).

With prevalence rates as high as 7.3%, generalized anxiety disorder (GAD) is one of the most common anxiety disorders among older adults and may well be one of the most common psychological disorders in this age group (1-to-12 month prevalence rate; Beekman et al., 1998; Flint, 2005; Gum, King-Kallimanis & Kohn, 2009; Riedel-Heller, Busse & Angermeyer, 2006). GAD is characterized by excessive worry that is hard to control, and is accompanied by symptoms such as restlessness, trouble sleeping or concentrating, irritability, fatigue and muscle tension. Research on the psychopathological features of GAD is well-developed and it is thought that the phenomenology of this disorder is similar across age groups with comparable levels of worry in younger and older adults (Alwahhabi, 2003). What seems to differ across age groups is the worry content, which reflects content consistent with common age-related changes (Diefenbach et al., 2001). It is estimated that approximately half of older GAD patients develop the disorder in the later part of life (Flint, 2005) and that the late-life risk factors include chronic health problems, poor self-perceived health and functional limitations (Vink, Aartsen, & Shoevers, 2008).

Many factors can complicate the assessment and diagnosis of late life anxiety. First, anxiety can coexist with depression within all age groups and the symptoms of these disorders often overlap (Kessler et al., 2005). For example, as is the case with depression, anxiety can provoke changes in cognitive functioning such as an inability to concentrate, remember or make decisions (Erber, 2010). Further, it has been noted that anxiety and depression are highly intercorrelated in older adults (Shamoian, 1991) and that differential diagnosis may be particularly challenging. Studies suggest that depression is the most common comorbid disorder among those with an anxiety disorder, with as many as 38 to 46% of older adults who meet criteria for an anxiety disorder also meeting criteria for a mood disorder (Beekman et al., 1998; Flint, 1994; Lenze et al., 2000).

Another factor that complicates the assessment of older adults is the fact that anxiety and physical illnesses in later life may produce symptoms that are difficult to distinguish from one another. With studies suggesting that between 80 and 86% of adults aged 65 and older have at least one significant medical condition (Dawson, Kline, Wiancko, & Wells, 1986; Haley, 1996; Naughton, Bennett, & Feely, 2006), many symptoms of anxiety may be overlooked and wrongly attributed to a medical illness. Consequently, older adults may be more likely to attribute physical symptoms related to anxiety (including muscle tension, hypervigilance and difficulties related to sleep) to a medical problem than to anxiety (Gurian & Miner, 1991). In turn, many physical conditions such as cardiovascular disease, respiratory disease, hyperthyroidism and pulmonary difficulties can mimic or cause anxiety symptoms, making it difficult to establish the underlying cause of these symptoms (Alwahhabi, 2003; Kogan, Edelstein & McKee, 2000). Additionally, medication side effect can also imitate psychological symptoms. For example, medication such as antidepressants or antihypertensives may produce anxiety-like symptoms such as insomnia or decreased concentration. It is therefore essential, when assessing an older patient, to obtain

enough information to make an appropriate diagnosis and to rule out possible differential diagnoses.

Anxiety disorders are also frequently comorbid with cognitive decline and dementia among older adults (Seignourel, Kunik, Snow, Wilson, & Stanley, 2008). Longitudinal research has demonstrated that anxiety in later life may increase the risk of cognitive decline over and above the risk associated with increasing age (Sinoff & Werner, 2003). However, the association between anxiety and cognitive decline may be bidirectional with chronic anxiety causing cognitive impairment or with anxiety developing after cognitive impairment, perhaps in response to the awareness that cognitive abilities are declining(Wolizky-Taylor, Castriotta, Lenze, Stanley & Craske, 2010). In addition, it might be difficult, particularly in the presence of cognitive decline, to distinguish whether symptoms such as agitation or restlessness are symptoms of dementia or anxiety, For example, studies have found that caregivers of patients with dementia often observed behavioral manifestations of anxiety such as restlessness and pacing (e.g. Fisher, Goy, Swinger & Szymanski, 1994). It is difficult to determine if the behaviors are evidence of an anxiety disorder or the agitation typically found in dementia. Cognitive decline is an important factor to consider when assessing anxiety in this age group because it may affect the presentation of the symptoms as well as the ability to communicate them to a clinician. Finally, older adults or clinicians may also misattribute anxiety symptoms such as fatigue and difficulty concentrating to normal aging processes.

In both clinical and research contexts, self-report measures appear to be the dominant method to assess the presence of anxiety (Alwahhabi, 2003; Dennis et al., 2007). However, because relatively little is known about the experience and presentation of anxiety in older adults, it is not clear if the current anxiety instruments accurately assess the unique characteristics of anxiety in older adults. For example, Lawton, Kleban & Dean (1993) found that older adults were

more likely to experience anxiety directly (e.g., feeling fearful or scared) than younger adults who experienced symptoms that were a mix of guilt and anxiety (e.g. ashamed, guilty, feeling like you did something wrong). Similarly, Kogan and his colleagues (2000) found that older adults reported several fears (e.g., being a burden for others) that were not found in fear surveys designed for use with adult samples. As older adults experience anxious symptoms in a different way than younger adults do, and as there are many factors that can complicate the use of self-report measures with older adults, it is critical that the anxiety measures used with older adults be demonstrated to be scientifically sound with this population. Due to the small number of studies in this area, and their inconsistent findings, there is a need for more research examining the psychometric properties of anxiety scales used with older adults.

Current Theories of Measurement

Psychometric theory has several different measurement approaches within its general framework, including classical test theory (CTT) and item response theory (IRT). CTT has received the most attention in the literature, as it is simple to employ and appears to have been the most commonly used theory by measurement researchers for many years. The fundamental assumption of CTT is that a respondent's observed score on a scale or test represents his or her "true score" plus random error. True score can be defined as the mean of the theoretical distribution of test scores that would be obtained from repeated individual testing of the same person with the same test. Error consists of random unsystematic deviations from true score that occur in each testing occasion. Because error is random it varies in every test administration and consequently so does the observed score. CTT estimates of traits are test-dependent and every test (and test administration) has different psychometric properties. On the other hand, IRT approaches the measurement problem in a manner that is quite different from CTT. The fundamental assumption underlying IRT is that every respondent has some "true" location on a

continuous latent dimension (often called theta). This location is assumed to represent the probabilistic influence of a person's response to any item or set of items that relates to the trait that theta represents. IRT models theta by using mathematical equations that relate responses patterns to a set of items, the psychometric properties of these items, and the knowledge of how item properties influence responses.

Due to their length, and to both the time and procedural separation between rating, scoring and interpreting, psychometric measures have not always been widely accepted in medicine. In an effort to create clinically relevant measurement procedures, physicians and other health researchers have utilized a theoretical approach referred to as clinimetrics (Feinstein, 1987) and more recently communimetrics (Lyons, 2009). Whereas psychometric measures are mainly designed for research purposes, clinimetrics and communimetrics assessment tools are designed to be congruent with the clinical process. Assessment tools created by both these approaches are therefore designed so they can operate at the item level. In other words, because these approaches are action oriented, the items that are included in the measures have meaningful and direct links to what happens next in the services. The individual items are selected to guide decision making, as they indicate what level of service effort is required (i.e., 0 = No evidence, no need for action, 1 = Watching waiting or monitoring only required and 2 = Action is needed). Consequently, they are designed to be accessible to service providers, consumers and policy makers.

Psychometric Properties of Psychological Instruments

For many assessment instruments, strong evidence to support their clinical usefulness is lacking, and many commonly used assessment methods and measures are not supported by scientific evidence (Hunsley & Mash, 2008b). There seems to be a belief in the intrinsic scientific value of many commonly used assessment measure, which may reflect a lack of understanding regarding the evidence needed to ascertain the value of psychological measures. Simply put,

researchers and clinicians should use assessment instruments that show strong scientific evidence and that are appropriate for both the population and the task at hand. More precisely, for any measure, there needs to be reliability and validity evidence for the population being studied, as well as age-appropriate norms. The following section addresses these characteristics in more details.

Reliability. Reliability is an important psychometric property that needs to be taken into account when evaluating the usefulness of a specific measure (Hunsley & Mash, 2008b). Based on psychometric theory, reliability refers to the consistency of measurements and can be defined as the "degree to which test scores are free from errors of measurement" (American Psychological Association, 1985, p.19). A measure must be consistent within itself (internal reliability), consistent over time (test-retest reliability) and consistent if used by another rater (inter-rater reliability). According to classical psychometric theory, part of the variance of a set of scores is explained by the characteristic measured in the test (true score) but other influences, like sampling or measurement error (error score), can also explain a part of this variance (Graham, 2006). A measure is not considered to be reliable if scores are subject to large unsystematic fluctuations due to measurement error.

Reliability estimates demonstrate the proportion of variance in scores that is explained by the measure itself and therefore express the degree of consistency in the measurement of test scores. Thus, a reliability estimate of 0.00 represents the absence of reliability and a value of 1.00 a perfect reliability. Low reliability signifies that meaningful sources of errors are affecting the measure and that it cannot be considered to be consistent. According to Hunsley and Mash (2008b), a measure is considered to have adequate reliability when estimated values range from 0.70 to 0.79, good reliability between 0.80 and 0.89 and excellent reliability when exceeding

0.89. Examining the reliability of anxiety instruments commonly used with older adults is the central focus of this dissertation.

Although this dissertation is based the CTT approach, it is worth examining the role of reliability in other measurement theories. In terms of IRT, the goal of measurement is to reliably locate a person on the continuum relative to other possible individuals. In IRT the concept of score fidelity or reliability is replaced by the concepts of item and test information. In contrast, given the considerations underlying a communimetric tool, one would not expect different items to necessarily correlate with each other. Thus, internal consistency *per se*, is not relevant to the evaluation of the individual items of a measure under this theory. There might be items that are related statistically, but all such items would be included because they may have different action implications. The decision to drop or add an item is understood from the value of the relation to future clinical actions, not to the statistically inter-relation among items.

Validity. When applied to clinical assessment, validity most commonly refers to the degree to which a measure reflects the phenomenon it is supposed to measure. In other words, validity is the degree to which real-life changes on a dimension of an attribute in an individual produce changes in a measure of that attribute dimension (Borsboom, Mellenbergh & van Heerden, 2004; Messick, 1995; Nunally & Bernstein, 1994). A measure must include items that are representative of all aspects of the underlying psychological construct that the measure is designed to assess (content validity), provide data consistent with other constructs associated with the construct of interest (concurrent validity) and provide a pure measure of the construct that is not contaminated by other psychological constructs (discriminant validity) (Hunsley & Lee, 2010). Because different aspects of the measured construct can vary across populations, a measure cannot be considered as simply valid or invalid. The fact that a measure has supporting validity evidence for a specific purpose, with a specific group, does not mean it is valid for other purposes or groups. Thus, when deciding whether a measure is appropriate to be used with a

client, the psychologist needs to determine if there is validity based on research with members of the same population as that of the client (Hunsley & Lee, 2010).

Norms. Norms or specific criterion-related cut-offs scores are necessary if one wants to interpret the results of a measure in a meaningful way (AERA et al., 1999). A norm provides an indicator of the average typical performance of a specific group and the spread of the scores above and below the mean. Without norms, the results on a measure are not meaningful and cannot be interpreted (Hunsley & Mash, 2008b). For example, if an older adult was to obtain a score of 20 on an anxiety measure, it would provide no useful information unless the range of scores obtained by other older adults is known. A critical aspect of an instrument's norms is the quality of the normative sample. It is all too common for norms to be based on a convenience sample, such as undergraduate students or hospital inpatients. However, when it comes to evaluating older adults, it is important for the person being evaluated to be comparable to the normative group on variables such as age. As few measures are created for older adults, it is unlikely that age-appropriate norms are available for interpreting a score, thus making the interpretation of the results quite challenging and potentially very problematic. According to Hunsley and Mash (2008b), in order to be considered as having adequate norms, an instrument needs measures of central tendency and distribution for the total score based on a large, relevant sample.

Reliability Induction and Reliability Generalization

The reliability (i.e. internal consistency, inter-rater, test-retest) of generated scores is critical to the determination of a measure's merit, as a measure cannot be scientifically sound without evidence of reliability. It is important to note that a measure in itself is not reliable or unreliable but that reliability is a property of the scores obtained by a certain sample on the measure (Henson, 2001; Hunsley & Mash, 2008b; Rousse, 2007). In other words, a measure can produce reliable scores in a study and then unreliable scores in another study with different participants. Unfortunately, even with the growing attention that has been placed on clarifying the

nature of reliability coefficients, it is still common for some researchers to act as if reliability coefficients from other studies or test manuals necessarily apply to their own research samples (Rousse, 2007; Vasaar & Bradley, 2010). Research in the last few years has shown that not only do researchers often fail to report reliability coefficients for their own data but that many rely upon the use of reliability induction (Rousse, 2007; Vasaar & Bradley, 2010). Reliability induction consists of reporting reliability coefficients from previous samples or test manuals instead of calculating reliability estimates from the data provided by the study sample(s).

Because of the sample-specific nature of reliability, Vacha-Haase (1998) proposed a meta-analytic method known as reliability generalization (RG) in order to estimate the mean reliability score of a measure based on the obtained reliability coefficients of different studies. By reviewing the obtained score reliabilities of different studies for a specific measure, the RG can (a) provide the mean score reliability associated with the measure, (b) evaluate the variability in score reliability across samples for the measure and (c) determine which variables can predict or explain the variations in score reliability for a specific measure. This technique has been shown to be a powerful tool, and an increasing number of studies have used the RG meta-analysis in the last few years to examine the reliability of instruments designed to assess a variety of psychological construct (e.g., Lopez-Pina, Sanchez-Meca &Rosa-Alcazar, 2009; Vassar & Bradley, 2010).

The Classification of Psychological Measures as Evidence-Based

Evaluating whether a specific assessment measure can be considered as evidence-based can be an arduous task as a measure can produce reliable scores in one study and unreliable scores in another. For this reason, psychometricians and test developers are reluctant to establish precise standards for psychometric properties that an instrument must meet in order to be considered useful for different assessment purposes (Streiner & Norman, 2007). This is of little

help to clinicians who need to decide whether an instrument is good enough for the task at hand and who, without specific guidelines, might turn to frequently used measures that have little or no psychometric support. Although meta-analytic estimates for reliability and validity are ideal to gather information on the likely psychometric properties of a measure when used for a specific purpose and with specific populations, obtaining these estimates is extremely time-consuming and labour-intensive. Fortunately, recent steps have been taken towards categorization schemes that operationalize the criteria necessary to designate a psychological instrument as evidence-based. At this point, three such approaches have been described in the literature.

The first comprehensive effort to develop and apply criteria for evaluating clinical instruments was that of Bickman, Nurcombe, Townsend, Belle, Schut, and Karver (1999). As part of the reforms of the Australian government to improve the field of mental health services, the authors completed a review of measures used in child and adolescent services. Databases searches of relevant articles, chapters and books published between 1990 and 1998 yielded a list of 188 measures. Information for each measure was obtained and reviewed in order to rate them on the basis of 29 criteria (see Table 1) related to psychometric qualities, cultural sensitivity, developmental sensitivity, feasibility and cost. Most criteria involved simple yes or no ratings or similar ratings, but the researchers also developed specific criteria for determining specific levels of reliability and validity ranging between unacceptable and highly acceptable.

Table 1 Bickman et al. (1999) Criteria for rating outcome measures in child and adolescent mental health services

Cultural sensitivity	Evidence of bias Cultural norms Used in country of assessment
Suitability	Cost Time required Training required Experience required to administer

	Reading level Computer software available
Reliability	Test-retest reliability within a 2-week period Test-retest coefficient Internal consistency Cross-informant agreement Inter-rater agreement
Content validity	Theory-based items Expert judgment used in item evaluation Respondent feedback used in item evaluation Factor analytic findings
Construct validity	Convergent validity Divergent validity Social desirability Group differences Sensitivity/specificity
Norms	Current norms available Number of normative samples
Developmental sensitivity	Number of forms for specific age groups Age effects Norms across ages and age groups

After this initial coding, the authors created 18 evaluation requirements that were most relevant for determining the best measures for clinical purposes. Among these 18, five requirements were chosen as being most important: the instrument took less than ten minutes to complete, required less than 2 hours of training for the person administering and interpreting it, required no specialized background for using the instrument, had evidence of convergent validity coefficients greater than 0.5 and had known groups validity. Measures meeting the 5 minimum requirements could be judged as recommended for clinical use. The strength of this approach resides in the use of a systematic search to generate a list of commonly used measures to be evaluated, the use of appropriate evaluation criteria and the identification of key elements. However, the considerable volume of work involved in the search and the coding of criteria makes this approach almost impossible for clinicians to implement when selecting a measure for their assessment purposes.

Hunsley and Mash (2008a) recently elaborated a similar approach in which explicit criteria were used to evaluate the appropriateness of assessment instrument. It did, however, differ from the work of Bickman et al. (1999) as it was much broader and allowed the evaluation of measures used to assess a range of mental heath problems (e.g., anxiety disorders, mood disorders, couple distress, personality disorders) and different populations (e.g., children, adolescents, adults, older adults, couples). Subject area experts chose assessment measures that had either shown clinical utility or that were likely to be clinically useful. All chosen instruments were evaluated against the criteria developed by the authors. These criteria were based on a "good enough" principle, meaning that instead of focusing on the ideal criteria for a scientifically supported measure, the authors created criteria that indicated the minimum sufficient evidence to justify their use for clinical purposes.

The rating focused on nine criteria: norms, internal consistency, inter-rater reliability, test-retest reliability, content validity, construct validity, validity generalization, sensitivity to treatment change and clinical utility. For each criterion, the invited experts evaluated the available evidence and gave a rating of less than adequate, adequate, good, excellent, unavailable or not applicable for each criterion (see Table 2). What constitutes adequate, good and excellent varied for each criterion, but, in general, a rating of adequate indicated that the minimal level of scientific rigor was met, good indicated solid scientific support and excellent indicated close to ideal supporting evidence. On the basis of their evaluation, experts indicated which instruments were the best available measure for a specific purpose and disorder and could therefore be recommended. The strength of this approach resides in the use of subject area experts, the presentation of specific requirements of psychometric evidence and the broad range of clinical conditions reviewed. The limitation resides in the fact that some solid instruments might have been overlooked as they were chosen by the experts who might have preferred some instruments

over others. Additionally, the ratings for some dimensions (e.g., clinical utility) were based on more limited evidence than was available for other dimensions (e.g., reliability, validity).

Table 2 Hunsley & Mash (2008a) Criteria for Rating Assessment Instruments

Norms:
Adequate: Measures of central tendency and distribution for the total score (and subscores if relevant) based on a large, relevant, clinical sample are available.
Good: Measures of central tendency and distribution for the total score (and subscores if relevant) based on several, large, relevant samples (must include data from both clinical and nonclinical samples) are available.
Excellent: Measures of central tendency and distribution for the total score (and subscores if relevant) based on one or more large, representative samples (must include data from both clinical and nonclinical samples) are available.
Internal consistency:
Adequate: Preponderance of evidence indicates alpha values of .70-.79.
Good: Preponderance of evidence indicates alpha values of .80-.89.
Excellent: Preponderance of evidence indicates alpha values ≥ .90.
Inter-rater reliability:
Adequate: Preponderance of evidence indicates *k* values of .60-.74; the preponderance of evidence indicates Pearson correlation or intraclass correlation values of .70-.79.
Good: Preponderance of evidence indicates *k* values of .75-.84; the preponderance of evidence indicates Pearson correlation or intraclass correlation values of .80-.89.
Excellent: Preponderance of evidence indicates *k* values of ≥ .85; the preponderance of evidence indicates Pearson correlation or intraclass correlation values of ≥ .90.
Test-retest reliability:
Adequate: Preponderance of evidence indicates test-retest correlations of at least .70 over a period of several days to several weeks.
Good: Preponderance of evidence indicates test-retest correlations of at least .70 over a period of several months.
Excellent: Preponderance of evidence indicates test-retest correlations of at least 0.70 over a period of a year or longer.
Content validity:
Adequate: The test developers clearly defined the domain of the construct being assessed and ensured that selected items were representative of the entire set of facets included in the domain.
Good: In addition to the criteria used for an Adequate rating, all elements of the instrument (e.g., instruction, items) were evaluated by judges (e.g., by experts, by pilot research participants).
Excellent: In addition to the criteria for a Good rating, multiple group of judges were employed and quantitative ratings were used by the judges.
Construct validity:
Adequate: Some independently replicated evidence of construct validity (e.g., predictive validity, concurrent validity, convergent and discriminant validity).
Good: Preponderance of independently replicated evidence, across multiple types of validity

(e.g., predictive validity, concurrent validity, convergent and discriminant validity).
Excellent: In addition to the criteria used for a Good rating, evidence of incremental validity with respect to other clinical data.
Validity generalization:
Adequate: Some evidence supports the use of this instrument with either a) more than one specific group (based on sociodemographic characteristics such as age, gender and ethnicity or b) in multiple settings (e.g., home, school, primary care setting, inpatient setting).
Good: Preponderance of evidence supports the use of this instrument with more than one specific group (based on sociodemographic characteristics such as age, gender and ethnicity) or b) in multiple settings (e.g., home, school, primary care setting, inpatient setting).
Excellent: Preponderance of evidence supports the use of this instrument with more than one specific group (based on sociodemographic characteristics such as age gender and ethnicity) and across multiple contexts (e.g., home, school, primary care setting, inpatient setting).
Treatment sensitivity:
Adequate: Some evidence of sensitivity to change over the course of treatment.
Good: Preponderance of independently replicated evidence sensitivity to change over the course of treatment.
Excellent: In addition to the criteria used for a Good rating, evidence of sensitivity to change across different types of treatments.
Clinical utility:
Adequate: Taking into account practical considerations (e.g., costs, ease of administration, availability of administration and scoring instructions, duration of assessment, availability of relevant cutoffs scores, acceptability to patients), the resulting assessment data are likely to be clinically useful.
Good: In addition to the criteria used for an Adequate rating, there is some published evidence that the use of the resulting assessment data confers a demonstrable clinical benefit (e.g., better treatment outcome, lower treatment attrition rates, greater patient satisfaction with services).
Excellent: In addition to the criteria for an Adequate rating, there is independently replicated published evidence that the use of the resulting assessment data confers a demonstrable clinical benefit.

A different approach to classifying measures was taken in a recent task force of the American Psychological Association's Society of Pediatric Psychology (Cohen et al., 2008). The goal of the task force was to systematically evaluate assessment instruments used in the field of child health care and provide a guide for identifying the most appropriate instruments for research and clinical use. The task force and a set of experts identified measures that fell into one of eight construct areas of broad interest for clinical assessment purposes: quality of life, family functioning, psychosocial functioning and psychopathology, social support and peer relations,

treatment adherence, pain, stress and coping and cognitive functioning. Work groups for each area of interest selected and reviewed the most widely used and researched instruments by using a combination of the results of a survey posted on the Society's listserve, the current literature, and their own expertise. The criteria developed to determine if a measure was evidence-based were based on those used to identify evidence-based treatments. Depending on the available empirical support, each instrument could be evaluated as Promising, Approaching Well-Established, and Well-Established (see Table 3). The work groups were also asked to assess the clinical and research utility of the measure and to identify which measures could be used with various ethnic and linguistic minorities.

Because the list of possible measures was generated by professionals working in the field of pediatric psychology, the pool contained many measures that were scientifically sound and clinically useful. However, the lack of systematic literature review strategies might result in some relevant measures being overlooked. As with Hunsley and Mash (2008a), using subject area experts was both a strength and a limitation as it ensured that knowledgeable individuals evaluated the instruments but also might have created subjective biases. Additionally, although using empirically treatment-like criteria provided some form of face validity, given that assessment can be undertaken for a range of purposes, the criteria might not be specific enough to provide a meaningful evaluation of the many purposes for which a measure might be used.

Table 3 Cohen et al. (2008) Criteria for Evidence-Based Assessment

Well-Established assessment
▫Measure presented in at least two peer-reviewed articles by different investigators
▫Sufficient detail about the measure to allow critical evaluation and replication
▫Detailed information indicating good validity and reliability in at least one peer-reviewed article
Approaching Well-Established assessment
▫Measure presented in at least two peer-reviewed articles which can be by the same investigator

▫Sufficient detail about the measure to allow critical evaluation and replication ▫Validity and reliability presented in vague terms or moderate values
Promising assessment ▫Measure presented in at least one peer-reviewed article ▫Sufficient detail about the measure allow critical evaluation and replication ▫Validity and reliability presented in vague terms or moderate values

Despite a common goal to establish standardized criteria for determining the extent of evidence base supporting assessment instruments, all three approaches presented significant differences in term of the criteria used to evaluate the instruments, the search strategies used as well as the manner in which criteria was applied. This does raise questions regarding the utility and comparability of each approach. Do the three approaches yield similar results and, most importantly, do they provide conclusions similar to what would be obtained by meta-analytic investigations?

Summary

The current era of evidence-based practice depends on the use of scientifically sound assessment methods and instruments. As clinicians rely on the accuracy of assessment instruments for diagnosis and case conceptualization purposes, as well as decisions regarding the choice and efficacy of interventions, more attention has been recently paid to the field of evidence-based assessment. With the increasing proportion of older adults in the population, and the frequency with which they experience anxiety, it is surprising that so little is known about the evidence-based assessment of anxiety in older adults. It seems that, due to a lack of assessment instruments created specifically for older adults, it is at times necessary to use measures developed for younger populations of adults when working with older populations. However, given the fact that older adults experience anxiety differently than younger adults, and as there are many factors that can complicate the use of self-report measures with older adults, it is critical

that the anxiety measures with older adults be demonstrated to be scientifically sound. Due to the small number of studies in this area, there is a need to further investigate the psychometric properties of anxiety scales when used with an older population.

This dissertation aims to add to the field of evidence-based assessment in the older adult population by identifying the instruments that are the best suited to assess the presence of anxiety in older adults. The first study identifies the measures most commonly used to assess anxiety in older adults through a systematic review of the literature. The psychometric properties of each measure (reliability, validity, norms, clinical utility) are thoroughly reviewed in order to evaluate whether they are appropriate for use with older adults. The second study, a reliability generalization analysis, further examines the reliability of the most commonly used anxiety measures identified in the first study. The purpose of this meta-analytic research is to estimate the average reliability of each anxiety measures, to examine how reliability estimates are influenced by factors such as participant age and sample characteristics, and to determine whether the most commonly used anxiety measures are likely to provide reliable scores. With the meta-analytically based reliability data gathered in the second study, the third study will evaluate two of the categorization schemes that operationalize criteria necessary to designate a psychological instrument as evidence based. This will allow for an evaluation of the efficacy of each approach and will demonstrate whether the results provide conclusions similar to what was obtained in the meta-analytic investigation conducted in the second study.

Assessment of Anxiety in Older Adults: A Systematic Review of Commonly Used Measures*

Zoé Therrien John Hunsley

University of Ottawa, Canada

* This study was published in Aging and Mental Health: Therrien, Z., & Hunsley, J. (2012). Assessment of anxiety in older adults: A systematic review of commonly used measures. *Aging and Mental Health, 16,* 1-16.

Abstract

Objectives: The authors set out to systematically review the research literature in order to identify the anxiety measures most commonly used in the assessment of older adults. Once identified, the literature was reviewed to determine the extent to which these instruments had age-relevant norms and psychometric data supporting their use with older adults

Method: Literature searches were conducted in PsycINFO and PubMed to identify research articles in which anxiety measures were completed by older adults. After screening for suitability, a total of 213 articles were reviewed to determine the most commonly used anxiety measures with older adults, to examine the psychometric properties of these instruments, and to evaluate whether the instruments are appropriate for use with older adults

Results: A total of 91 different anxiety measures were used in the 213 included articles. Twelve anxiety measures were most commonly used in the literature and of those three were specifically developed for older adults.

Conclusions: Of the most commonly used measures, the majority lacked sufficient evidence to warrant their use with older adults. Based on psychometric evidence, three measures (Beck Anxiety Inventory, Penn State Worry Questionnaire, and Geriatric Mental Status Examination) showed psychometric properties sufficient to justify the use of these instruments when assessing anxiety in older adults. In addition, two measures developed specifically for older adults (Worry Scale and Geriatric Anxiety Inventory) were also found to be appropriate for use with older adults.

Keywords: older adults, anxiety, assessment, psychometric properties, systematic review

Assessment of Anxiety in Older Adults: A Systematic Review of Commonly Used Measures

Over the last decade, there has been a growing emphasis on the importance of evidence-based mental health practice (American Psychological Association Presidential Task Force on Evidence Based Practice, 2006; Institute of Medicine, 2001). According to the Institute of Medicine (2001), evidence-based practice in heath care services integrates information derived from the best research evidence, clinical expertise and the patient's values when contemplating health care services for a patient. This framework is based on the acknowledgement that positive outcomes are less likely if the chosen service does not have a research base that shows the potential to improve the client's functioning (Forman, Fagley, Steiner, & Schneider, 2009). The evidence-based practice movement has been endorsed through reports released by several scientific organizations in the field of health and mental health.

Much of the efforts to evaluate and disseminate evidence-based mental health practices have focused on intervention services. For example, a growing body of research has concentrated on the identification of efficacious psychological interventions (e.g. Ayers, Sorrell, Thorp, & Wetherell, 2007) and the establishment of guidelines for determining whether an intervention can be considered evidence-based (e.g., Chambless & Hollon, 1998). As assessment is a necessary component of health care services that should inform treatment selection and implementation, it seems incongruous that so little effort has been put towards identifying and promoting evidence-based assessment (Hunsley & Mash, 2007). Indeed, without scientifically sound assessment, clinicians cannot clearly evaluate a patient's level of functioning; without this information, the development of solid case formulations (including diagnostic considerations) is not possible, and without high quality case formulations, it becomes difficult to make informed treatment choices. Moreover, in research contexts, assessment measures are used to select participants and evaluate

treatment, thus they are critical for developing evidence-based treatment (Cohen et al., 2008). In sum, whether assessment measures are used for diagnostic purposes, to select research participants, to establish case conceptualization, to inform the choice of a treatment plan, or to monitor treatment outcome, the choice of measures is central to the quality of services provided. Fortunately, some recent efforts have been made towards the development of general guidelines to be used in selecting the best instruments. These efforts emphasize, in particular, the importance of solid psychometric properties, appropriate norms, and evidence of clinical utility (Holmbeck & Devine, 2009; Hunsley & Mash, 2008).

Although research on evidence-based assessment is increasing, research on the assessment of older adults is much less developed (Ayers et al., 2007). With the growing number of older adults in the general population, there is also a concomitant rise in the number of older adults who require mental health services. As diagnosis and treatment selection is informed by assessment data, it is necessary to have measures that are appropriate for an older population, but the lack of research evidence for the psychometric quality of many of these measures makes it challenging to choose an appropriate measure for use with older adults (Edelstein et al., 2001).

The empirical literature on anxiety prevalence suggests that it has become a widespread problem in late life. With 2-to-12-month prevalence estimates ranging from 1.2 to 15% in community samples of older adults (Bryant, Jackson & Ames, 2009; Wolitsky-Taylor, Castriotta, Lenze, Stanley, & Craske, 2010), and from 1 to 28% in clinical samples of older adults (Bryant et al., 2009), it is more common than depression. Furthermore, among hospitalized geriatric patients, the prevalence of anxiety disorders in the preceding 3 days is estimated to be as high as 43%, and is known to remain high after discharge (Kvaal, Macijauskiene, Engedal, & Laake, 2001). Moreover, the prevalence of older adults with anxiety symptoms that do not meet criteria for an anxiety disorder has been found to range from 15 to 52.3% in community samples and 15

to 56% in clinical samples (Bryant et al., 2009). There is significant impairment and a lower level of quality of life among anxious individuals (Mendlowicz & Stein, 2000) and, according to de Beurs and his colleagues (1999), older adults who present some anxiety symptoms are as negatively affected in their quality of life as those who meet criteria for an anxiety disorder.

Unfortunately research on anxiety disorders in older adults has not grown at the same rate as research on anxiety disorders in younger populations (Dennis, Boddington, & Funnell, 2007). One consequence of this is that relatively little is known about the evidence-based assessment of anxiety in older adults. In both clinical and research contexts, self-report appears to be the dominant assessment method for gathering information on the experience of anxiety (Alwahhabi, 2003; Dennis et al., 2007). For self-report measures to provide valid information, information on additional factors such as the frequent comorbidity of mental and physical health problems (Cully et al., 2006; Wolitsky-Taylor et al., 2010) and the use of multiple medications in older populations must be obtained and incorporated into assessment decisions.

The assessment and diagnosis of late life anxiety is especially challenging, as symptoms of anxiety can be confused with some aspects of the normal aging process (Lenze & Wetherell, 2009) as well as with medical conditions and comorbid mental disorders (Kogan, Edelstein, & McKee, 2000). It is well-established that anxiety and depression are frequently comorbid in younger adults (de Graaf, Bijl, Spijker, Beekman, & Vollebergh, 2003; Kessler, Berglund, Demler, Jin, & Walters, 2005). In the Epidemiological Catchment Area (ECA) study (Weissman et al., 1988) of people aged 18 to 54, 20% of individuals who received a diagnosis of any anxiety disorder in the past 6 months also received a diagnosis of some type of affective disorder (Regier, et al., 1988). With regard to the older adult population, one study using a community-based sample of Canadian adults aged 55 and older found that depression was the most common comorbid disorder among those with anxiety disorders, with 23% of those with an anxiety

disorder also meeting criteria for MDD (Cairney, Corna, Velhuizen, Hermann, & Streiner, 2008). Studies of depressed older adults also indicate that approximately half of these individuals meet criteria for an anxiety disorder (e.g., Beekman et al., 2000).

The detection of anxiety in older adults is also complicated by the high frequency of medical disorders in this age group. With studies suggesting that between 80 and 86% of adults aged 65 and older have at least one significant medical condition (Dawson, Kline, Wiancko, & Wells, 1986; Haley, 1996; Naughton, Bennett, & Feely, 2006), many symptoms of anxiety may be overlooked or wrongly attributed to a medical illness. Older adults may be more likely to attribute physical symptoms related to anxiety (including muscle tension, hypervigilance and difficulties related to sleep) to a medical problem than to anxiety (Gurian & Miner, 1991). In turn, many physical conditions such as cardiovascular disease, respiratory disease, hyperthyroidism and pulmonary difficulties can involve anxiety symptoms, making it difficult to establish the underlying cause of these symptoms (Alwahhabi, 2003; Kogan et al., 2000). Additionally anxiety symptoms can occur as side-effects of medication that is being used to treat a medical condition. In assessing elderly patients with anxiety complaints or unexplained physical symptoms, clinicians need to obtain enough information to make an appropriate diagnosis or, at the least, to describe possible differential diagnoses. Moreover, current anxiety diagnostic criteria and measures, most of which were developed originally for use with much younger persons, are weighted heavily with somatic items, making it difficult to distinguish between medical and psychological causes of anxiety in this population. Methods for assessing anxiety in older adults can be enhanced through consideration of unique aspects of anxiety in this population, including age-relevant aspects of physical and mental health status.

Anxiety disorders are also frequently comorbid with cognitive decline and dementia among the elderly (Seignourel, Kunik, Snow, Wilson, & Stanley, 2008). This comorbidity is

partially due to the fact that there may be a specific relation with anxiety and several types of cognitive impairments (Wolizky-Taylor et al., 2010). Cognitive decline is an important factor to consider when assessing anxiety in this age group because it may affect the presentation of the symptoms as well as the ability to communicate them to a clinician. What seems to be symptoms of anxiety in older adults (e.g. agitation) may in reality be the result of the challenges associated with memory impairment. Finally, older adults or clinicians may misattribute anxiety symptoms (e.g., fatigue, difficulty concentrating) to normal aging processes. It has also been suggested that the previous cohort of older adults might be less comfortable than current generations in discussing emotions and, therefore, more likely to minimize their symptoms (Pachana, 2008).

In sum, as older adults experience anxious symptoms in a different way than younger adults do, and as there are many factors that can complicate the use of self-report measures with older adults, it is critical that the anxiety measures used with older adults be demonstrated to be scientifically sound. In recent years, researchers have reviewed anxiety measures often used with older adults (e.g., Edelstein et al., 2008; Kogan, Edelstein & McKee, 2000). Although such reviews provide an important first step in evaluating the scientific merits of available instruments, they do have significant limitations. Most notably, neither Edelstein et al. (2008) nor Kogan et al. (2000) indicated the bases for selecting the instruments they chose to review. Furthermore, only limited psychometric information on selected instruments was presented in these reviews, thus making it difficult for readers to determine the scientific adequacy of these instruments for clinical or research purposes.

With this background in mind, we set out to systematically review the research literature in order to identify the anxiety measures most commonly used in the assessment of older adults. Once identified, our intent was to review the literature to determine the extent to which these instruments had age-relevant norms and psychometric data (reliability, convergent validity,

discriminant validity, and treatment sensitivity) supporting their use with older adults. By summarizing this research, we provide critical up-to-date information regarding the status of evidence-based instruments for assessing anxiety in older adults. As indicated, this information is crucial in developing and providing evidence-based mental health services to older adults.

Method

Study Eligibility and Search Strategy

The electronic databases PubMed and PsycINFO were searched for published journal articles that included some type of anxiety measure used in a population of older adults. Studies meeting the following criteria were selected: a) an empirical study, b) used an anxiety measure with at least one participant, and c) only included adults aged 65 years and above in the sample.

To ensure the broadest possible search of the databases, two separate searches combining different key words and study criteria were carried out in each database. The first search included the following terms: *anxiety*, *anxiety disorder, generalized anxiety disorder* AND *assessment*, *measurement*. The second search combined the following terms *anxiety*, *anxiety disorders*, *generalized anxiety disorder* AND *geriatric assessment*, *geriatric patients*, *geriatric psychiatry*, *geriatric*, *gerontology*. The key words were searched as article keywords, titles, and abstracts. The search was limited to older adults by selecting "Age 65 and older" in the Age Group section of the search. The search was restricted to articles published between January 1960 and August 2009, in English or French (as these were the languages understood by the authors). Articles were obtained through our university library network or, for the articles that were unavailable in print or electronically at our university, through inter-library loans. If an article was unavailable through both these methods, a search was conducted at the National Archives of Canada.

Search Results

The search yielded a total of 1,427 articles (785 in PubMed and 642 in PsycINFO). After reviewing the titles and abstracts, 592 unique articles were retained at this point for further examination to determine if they fully met our inclusion criteria. As part of this review, only studies presenting original data were included in our final set of retained articles. Articles were excluded because they did not have a sample that included only older adults (49.29%), did not use an anxiety measure (8.92%), did not have a sample of older adults and did not have an anxiety measure (7.04%), were not empirical studies (20.0%), did not use original data (6.29%), were not in French or English (5.16%), or were unavailable in print or electronic form within our university library network, through inter-library loans or through the National Archives of Canada (3.75%).

Data Coding

For all the 592 selected articles, the full article was obtained and reviewed to assess the fit with our inclusion criteria. After this detailed review, a total of 213 articles were retained. Articles were excluded because, after thorough review, they were found to not meet the inclusion criteria which required the articles to be empirical studies that employed an anxiety measure with adults aged 65 and older. The details of the 213 retained articles were coded to summarize key aspects of the study and the anxiety measures used in the study. For each anxiety measure, the rater noted if the authors indicated that the measure was appropriate for older adults, if it was designed for older adults, and if relevant norms are available. Participant variables (age range of the sample, the mean age and standard deviation of the sample, proportion of men and women included in the sample) and sample information (selection of participants, research setting) were also coded.

Results

General Characteristics of the Studies

Of the retained articles, close to half (45%) were in journals that specialize in research on older adults. The majority of articles meeting the inclusion criteria (83%) were published after 1997. The age of the participants in the retained articles ranged from 65 to 102 years, but the majority ranged between 65 and 75 years old. The samples reported in the majority of studies consisted of older adults recruited from within their communities (56%). With respect to study recruitment procedures, participants were also recruited from medical settings (38%), mental health settings (9%) and residential settings (11%). It is worth noting that the recruitment total does not add to 100% as some studies recruited participants in more than one setting (e.g., use of both a clinical and community sample in a study). In approximately half of the studies, the intent was to conduct research on normal, community-dwelling older adults; samples were also recruited because of having a specific medical disorder (26%), mental disorder (15%) or both (1%).

Most Commonly Used Measures of Anxiety

A total of 91 different anxiety measures were used in the 213 included articles. However, the majority of these measures were used in only one or two studies. Most (89%) of the anxiety measures were developed for use with younger adults but used with older aged samples; few measures (16%) were created with older adults' specific experiences or needs in mind. Of the measures created specifically for older adults, only one (Geriatric Mental State Examination, GMSE; Copeland et al., 1976) was commonly used in the studies we examined (14 times). Study authors rarely mentioned whether measures were appropriate for older adults (24%), and even fewer reported whether specific age-relevant norms were available to aid in the interpretation of the anxiety data (21%).

In our review of the scientific status of anxiety measures used with older adults, we focused on measures that had been used frequently in the literature, thus allowing for independent replication of findings and the availability of data from multiple types of samples. With this in mind, and in consideration of the measure frequency distribution we found in the 213 studies, we decided to concentrate on instruments that had been used in six or more studies. Using this criterion, the most commonly employed measures to evaluate anxiety in older adults were the State Trait Anxiety Inventory (STAI; Spielberger, Gorsush, Lushene, Vagg, & Jacobs, 1983; 35 studies), the Hospital Anxiety and Depression Scale (HADS; Zigmund & Snaith, 1983; 26 studies), the Geriatric Mental State Examination (GMSE; Copeland et al., 1976; 14 studies), the Hamilton Anxiety Rating Scale (HARS; Hamilton, 1959; 13 studies), the Goldberg Anxiety and Depression Scale (GADS; Goldberg et al., 1988; 8 studies), the General Health Questionnaire (GHQ, Goldberg, 1978; 8 studies), the Beck Anxiety Inventory (BAI; Beck, Epstein, Brown, & Steer, 1988; 8 studies), the Brief Symptom Inventory (BSI; Derogatis & Spencer, 1982; 7 studies), the Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, & Borkovec, 1990; 6 studies) and the Symptom Checklist-90-Revised (SCL-90-R; Derogatis, 1994; 6 studies).

Only one of the most commonly used measures (GMSE) was created specifically for older adults. In the past decade, considerable effort has been made to develop anxiety measures specifically suited for older adults. As such measures are relatively new and might not yet be frequently used in research on older adults, we decided to also include in our detailed review the most commonly employed measures created specifically for older adults. Accordingly, we relaxed our criteria to consider such measures. Based on our review of the literature, there were two instruments that have been used with sufficient frequency to allow for some replication of psychometric findings: the Worry Scale (WS; Wisocki, Handen, & Morse, 1986; 5 studies) and

the Geriatric Anxiety Inventory (GAI; Pachana, Byrne, Siddle, Kolowski, Harley, & Arnold, 2007; 4 studies).

Psychometric Evaluation of Measures

All these measures are standardized self-report questionnaires, clinician-administered rating scales, or standardized interviews that have been used to evaluate anxiety in a variety of settings. The measures and their psychometric properties are described in detail below. To give a fuller sense of the scientific adequacy of each measure, we describe the nature of the measures and provide information about the availability of age-appropriate norms. Additionally, we describe both reliability (internal consistency, test-retest reliability, and/or inter-rater reliability, as appropriate) and validity (convergent validity, discriminant validity, treatment sensitivity) evidence based on data from samples of older adults. The information regarding psychometric properties was obtained from the studies identified in the literature search as well as from the instrument manuals. [1]

Stait Trait Anxiety Inventory (STAI)

The STAI (Spielberger, Gorush, Lushene, Vagg, & Jacobs, 1983) is a 40-item self-report questionnaire, derived from the Minnesota Multiphasic Personality Inventory (MMPI), that is designed to measure and differentiate between anxiety as a trait and as a state. The two scales consist of 20 items answered on a four-item scale and provide a score ranging from 20 to 80, with higher scores indicating higher levels of anxiety. A cutoff score of 39/40 for both single scales is normally used to identify clinically significant symptoms of anxiety. However, optimal cutoff

[1] Some of the measures included in this study do not consist of anxiety-specific measures, but rather measures of general distress or psychopathology that include anxiety subscales (i.e., BSI, GADS, GHQ,HADS, SCL-90). For these measures, we provide information on the psychometric properties of the anxiety subscales and not the full measure. The only exception to this was the GHQ, as no study reported the psychometric properties of the anxiety/insomnia subscale.

scores for older adults have been found to range between 44 and 55 (Himmelfarb & Murrell, 1984; Kvall, Ulstein, Nordhus & Engedal, 2005). Means and standard deviations are available from the original authors for adults aged 50 to 69, but there are no normative data reported specifically for older adults.

The scale was developed for young and middle-aged adults, but further research has examined its psychometric properties with older adults. Good internal consistency has been shown for both the trait and state scales in older psychiatric outpatients (α = .92-.94 for the state version and α = .88-.90 for the trait version; Kabacoff, Segal, Hersen, & Van Hasselt, 1997; Stanley, Beck, & Zebb, 1996; Stanley, Novy, Bourland, Beck, & Averill, 2001) and in community dwelling older adults (α = .79-.90 for the trait version and 0.85 for the state version; Himmelfarb et al., 1984; Stanley et al., 1996). Test-retest reliability has been found to be good for the trait scale (*r* = .58-.84) and appropriately lower for the state scale (*r* = 51-.62; Stanley et al., 1996; Stanley et al., 2001). Unfortunately, there is only limited evidence of concurrent validity, as demonstrated by moderate correlations between the STAI-T and other measures of anxiety (*r* = .33-.57: Kabacoff et al., 1997; Stanley et al., 1996; Stanley et al., 2001); slightly lower correlations were found between the STAI-S and other measures of anxiety (*r* = .15-.52: Kabacoff et al., 1996; Stanley et al., 1996; Stanley et al., 2001). Both scales are substantially correlated with measures of depression (*r* = .41-.70; Stanley et al., 1996; Stanley et al., 2001), indicating only limited discriminant validity. There is also a concern that the STAI is lengthy and easily misinterpreted by older adults (Dennis et al., 2007).

Few studies identified in our review used the STAI to assess the effects of mental health treatments for older adults. In a sample of older adults receiving cognitive-behavior therapy (CBT) for GAD, the STAI was used before and after initiating treatment (5 to 20 weeks after the pretreatment assessment) (Stanley et al., 2001). Post-hoc paired comparisons of scores on the

STAI revealed a significant reduction in anxiety symptoms from pre- to post-treatment. Another study (Stanley et al., 2003) examined the efficacy of a CBT treatment relative to a minimal contact control in a sample of older adults with GAD. All participants completed measures of anxiety (STAI and HARS) and worry (WS, PSWQ) before and after initiating treatment. An analysis of simple effects demonstrated significant improvements on the STAI as well as on the HARS and PSWQ for the CBT group but not for the control group. Thus evidence from two studies suggests that the STAI can be sensitive to treatment effects. Although promising, evidence of the instrument's treatment sensitivity with other disorders and other treatments is needed. Overall, the studies examining the psychometric properties of the STAI have yielded mixed results and, therefore, it should be used with some caution when assessing anxiety in older adults.

Hospital Anxiety and Depression Scale (HADS)

The HADS (Zigmond & Snaith, 1983) is a 14-item self-report questionnaire developed to evaluate the presence and the severity of anxiety and depression in non-psychiatric outpatients. Items referring to symptoms that may have physical causes such as insomnia and dizziness were excluded from the scale during its development. It is, therefore, unlikely biased by comorbid medical conditions. The measure is divided into a 7-item anxiety subscale (HADS-A) and a 7-item depression subscale (HADS-D). The items are rated on a 4-point scale and summed to provide a score ranging from 0 to 21 for anxiety and for depression. There are no fixed cutoff scores for the HADS but, in their original study, Zigmond and Snaith (1983) recommended two cutoff scores: 7/8 for possible anxiety and depression and 10/11 for probable anxiety or depression. A cutoff score of 14/15 was later added for severe anxiety or depression, but no empirical data support this score (Snaith & Zigmond, 1994). The cutoff scores have not been validated with older adults but there is some evidence that use of the 7/8 cutoff score will

correctly identify the majority of anxious older adults (Haworth, Moniz-Cook, Clark, & Wang, 2007; Dennis et al., 2007). No norms are available for older adults.

The HADS was originally developed for medical outpatients aged between 16 and 65, but further research has since been conducted to validate its use with older adults. Even though not originally designed for use with psychiatric outpatients, the internal consistency of the HADS is high in samples of such individuals (α = .73-.80; Flint & Rifat, 2002; Wetherell, Birchler, Ramsdell & Unutzer, 2007). It has also been found to be high in samples of older medical inpatients (α = .75-.84; Bryant et al., 2009; Johnston, Pollard, & Hennessey, 2000, Yu, Lee, Woo, & Hui, 2007) and in community samples (α= .84-.85; Spinhoven et al., 1997). Although we found no study that reported evidence of test-retest reliability, the anxiety scale of the HADS has been found to show no significant changes when administered at a two-month interval in a population of inpatients in a geriatric hospital (Bryant et al., 2009). The HADS-A has demonstrated only limited evidence of concurrent validity, with moderate correlations with other psychological distress measures such as the BSI (r = .54; Wetherell & Aréan, 2007), the HARS (r=0.57; Dennis et al., 2007), and observer ratings of global anxiety (r = .28; Kenn, Wood, Kucyj, Wattis, & Cunane, 1987). It has shown moderate evidence of discriminant validity with depression measures (r = .26-.47), but moderate correlations between the anxiety and depression subscale (r=0.43-0.73) could result in high rates of misclassification (Davies, Burn, Mckenzie, Brotherwell & Wattis, 1993; Dennis et al., 2007; Flint et al., 2002; Johnson et al., 1995; Spinhoven et al., 1997).

A study conducted by Yu and his colleagues (2007) examined the effects of relaxation and exercise training on psychological outcomes in older patients with heart failure. Participants completed the HADS at baseline and at the 12th week of treatment. Analyses revealed that older adults who participated in the relaxation or exercise therapy showed lower levels of anxiety on

the HADS compared to the control sample. Although supportive of the instrument's treatment sensitivity, additional research assessing the effects of mental health treatments with the HADS are needed in the geriatric population.

Overall, although the HADS excludes somatic symptoms and shows high internal consistency, the correlations between the anxiety and depression scale suggests it may be most useful as an overall indicator of distress. These factors, when combined with the lack of evidence for clinical cut scores or norms relevant to older adults, suggest that this frequently used measure is not a good option for assessing anxiety in older adults (Bryant et al., 2009).

Geriatric Mental State Examination (GMSE)

The GMSE (Copeland et al., 1976) is a semi-structured clinical interview designed as a mental health assessment for older adults. The original version of the GMSE had 541 items, but it has since been shortened so that it can be administered in 20 to 50 minutes. The data can be analyzed by the computerized system AGECAT (Automated Geriatric Examination or Computer Assisted Taxomany) to obtain a suggested psychiatric diagnosis. Symptoms are grouped into eight syndrome clusters: organicity, schizophrenia and related paranoia, mania, depression, hypochondrias, phobias, obsessional and anxiety neurosis. A diagnostic confidence level is provided for each syndrome, ranging from 0 (no symptoms) to 5 (very severely affected). A level of three or more in a cluster represents a diagnostic case (Copeland, Dewey, & Griffiths-Jones, 1986).

The GMSE was developed and normed for older adults, and has become one of the most widely used comprehensive structured mental health assessments for older adults (Copeland et al., 2002). High inter-rater reliability for the general scale has been found in samples of community dwelling older adults, older medical patients, and older psychiatric inpatients due to its well-established procedures and structured approach to administering the questionnaire (*kappa*

= .73-.80; Ames, Flynn, Tuckwell, & Harrigan, 1994; Copeland et al., 1975; Turrina et al., 1991). Test-retest reliability has been somewhat inconsistent, with *r* values ranging from .49 to .75 (Copeland et al., 1976; Henderson, Duncan-Jones & Finlay-Jones, 1983). Evidence of concurrent validity in a variety of cultures has been shown with high correlations between the GMSE and DSM diagnostic criteria (*r* =.76-.78, Ames et al., 1994; Copeland et al., 1999). We were unable to find any study in which discriminant validity of the anxiety-related clusters was examined. There were no studies identified in our review in which the GMSE was used to assess the effects of mental health treatments for older adults. In sum the GMSE is a useful tool to assess the mental health of older adults in medical settings as it excludes the effects of physical illness. It has been subjected to many reliability and validation studies and is often used when examining the validity of other instruments used with older adults (Mottram, Wilson & Copeland, 2000). Although evidence regarding the validity of the anxiety-related clusters and the treatment sensitivity of the impairment is required, data on the correspondence with DSM diagnostic criteria suggest that it is likely to be useful in assessing clinically significant anxiety in older adults.

Hamilton Anxiety Rating Scale (HARS)

The HARS (Hamilton, 1959) is a 14-item clinician-administered rating scale developed to assess the severity of anxiety symptoms in adults. Seven of the items address psychic/cognitive anxiety and the remaining seven items address somatic anxiety. The items are rated on a 5-point scale and summed to provide a score ranging from 0 to 56. A score of 17 or less represents mild anxiety, a score between 18 to 24 mild to moderate anxiety, and a score of 25 and above moderate to severe anxiety. The cutoff scores have not been validated with older adults and there are no published norms for older adults (Kogan, Edelstein & McKee, 2000; Sheikh, 1991).

Although the HARS was developed for young and middle-aged adults, there is some recent support for its use with older adults. Adequate internal consistency has been shown in

samples of older adults diagnosed with generalized anxiety disorder (α = .77-.86; Beck, Stanley & Zebb, 1999; Diefenbach et al., 2001; Schuurmans et al. 2009). High inter-rater reliability has been shown with community samples and with older adults diagnosed with generalized anxiety disorder (*r* = .81-.95; Lenze et al., 2009; Stanley et al., 2009; Wetherell, Gatz & Craske, 2003). The HARS showed limited concurrent validity with the STAI-T (*r* = .23; Diefenbach et al., 2001) and the BAI (*r* = .47; Morin et al., 1999). The scale has been shown to differentiate older adults with generalized anxiety disorder from those with no anxiety disorders (Beck, Stanley & Zebb, 1996; Edelstein et al., 2008). However, the scale correlates considerably with the Hamilton Depression Scale (HDRS) in samples of older adults, raising concerns about its discriminant validity (*r* = .72-.92; Beck et al., 1996; Beck et al., 1999; Diefenbach et al., 2001). Importantly, the usefulness of the HARS with older adults has been questioned due to the heavy emphasis placed on somatic symptoms (e.g., tension) that are common experiences in aging individuals (Kogan et al., 2000; Skopp et al., 2006).

Several studies identified in our review used the HARS to assess the effects of psychological and pharmacological treatments for anxiety in older adults. In a randomized controlled trial comparing sertraline and CBT for the treatment of late-life anxiety, the HARS showed moderate to large effect sizes for the sertraline group and small to moderate effect sizes for the CBT group at both post-treatment and one year follow-up (Schuurmans et al., 2009). Similar results were found in a recent randomized controlled trial comparing the effect of escitalopram with a placebo (Lenze et al. 2009). However, studies examining the effect of CBT in samples of older adults with GAD have yielded mixed results. Two studies (Stanley et al., 2003; Wetherell et al, 2003) found that CBT participants had improved significantly on the HARS as well as on other anxiety measures, whereas another study (Stanley et al., 2009) found that changes on the HARS were not significantly different between the CBT and control group.

Overall, the HARS has shown high reliability in samples of older adults and some evidence of treatment sensitivity with psychopharmacological interventions. However, more studies are needed to establish its validity and treatment sensitivity across treatments in diverse groups of older adults. Given the lack of age-relevant norms and cut-off scores, and concerns about discriminant validity, this rating scale is not an optimal choice for use with older adults.

Goldberg Anxiety and Depression Scale (GADS)

The GADS (Goldberg et al., 1988) is an 18-item self-report questionnaire that measures symptoms of depression and anxiety experienced in the last month. The items are rated on yes (1) or no (0) answers and are summed to provide a score ranging from 0 to 9 for the depression subscale and for the anxiety subscale. According to Goldberg (1988), patients with anxiety scores of 5 or more or with depression scores of 2 or more have a 50% chance of a clinically important disturbance, and the probability of a significant disturbance increases substantially with higher scores. The anxiety cutoff score has not been validated with older adults and there are no normative data available for older populations.

The GADS was not developed for use with older adults and only a few studies have examined its psychometric properties with this population. The GADS showed good internal consistency in a sample of adults aged 18 to 79 where no substantial differences were noted when the alpha was calculated separately for different age groups ($\alpha = .82$; Christensen et al., 1999). Another study, conducted with older medical inpatients, reported good internal consistency ($\alpha = .82$; Huber, Mulligan, Mackinnon, Nebuloni-French & Michel, 1999). Although this shows initial evidence of reliability, no other published studies have examined reliability indices specifically in samples of older adults. Furthermore, we were unable to find any reports of test-retest reliability among samples of older adults. In the general population, the GADS correctly identified over 80% of adult patients with anxiety disorders as diagnosed by psychiatric assessment based on the

DSM-III criteria (Goldberg et al., 1988). However, when it comes to older adults, poor agreement has been found between the anxiety subscale of the GADS and other anxiety measures, with kappa values of -0.13 to 0.28 indicating poor concurrent validity (Kolowski, Smith, Pachana, & Dobson, 2008). Moderate correlations have been found between the GADS anxiety subscale and the GAI in a sample of older adults (Pachana et al., 2007). Finally, there is also evidence that the anxiety and depression subscale of the measure are highly correlated in both the general population and in samples of older adults, which suggests poor discriminant validity (Christensen et al., 1999; Huber et al., 1999; Kolowski et al., 2008). There were no studies identified in our review in which the GADS was used to assess the effects of mental health treatments for older adults. As a result there is no evidence with respect to the treatment sensitivity of the GADS. The inclusion of somatic symptoms (e.g., waking early, headaches) can overestimate the prevalence of anxiety and depression and result in classification errors. Overall, the psychometric evidence available thus far on the GADS is limited and provides little support for its use with older adults.

Beck Anxiety Inventory (BAI)

The BAI (Beck, Epstein, Brown, & Steer, 1988) is a 21 item self-report questionnaire designed to measure the severity of anxiety and to distinguish anxiety from depression. The items are rated on a 4-point scale and are summed to provide a score ranging from 0 to 63, with higher scores representing higher levels of anxiety. According to the manual (Beck & Steer, 1990), the score can be interpreted as follows: 0-9 (normal anxiety), 10-18 (mild to moderate anxiety), 19-29 (moderate to severe anxiety), and 30-63 (severe anxiety). Based on the information reported in the manual, it is unclear how those cutoffs were derived and there is no mention of whether different cut scores should be used with older populations. In subsequent research conducted with older adults, no single BAI cutoff proved to be optimal due to the tradeoffs between sensitivity and specificity (Kabacoff et al., 1997).

The BAI was developed and normed with samples of psychiatric adult outpatients. However, since its development, there have been several studies that evaluated its use with older populations. The internal consistency of data collected with the BAI is high in samples of older adult medical outpatients (α = .91-.92; Diefenbach, Tolin, Meunier, & Gilliam, 2009; Wetherell & Aréan, 1997), older adult psychiatric outpatients (α = .81-.93; Kabacoff et al, 1997; Wetherell & Gatz, 2005) and in community samples (α = .87-.89, Morin et al., 1999; Wetherell & al., 2005). The BAI showed adequate test-retest reliability (*r* = .64-.75) in samples of older adults (Beck et al., 1988; Diefenbach et al., 2009). Moderate correlations between the BAI and other anxiety measures show evidence of concurrent validity, with correlations ranging from .29 to .63 (Dennis et al., 2007; Diefenbach & al., 2009; Kabacoff et al., 1997; Wetherell & al., 2005). However, despite efforts in the development of the BAI to disentangle symptoms of anxiety and depression, relatively high correlations (*r* = .56-.65) between the BAI and depression measures show only limited evidence of discriminant validity (Wetherell et al., 1997).

In a study of older adults with GAD, participants were randomly assigned to a CBT group, a discussion group or a waiting list, and were assessed before and after treatment (Wetherell et al., 2003). At post-treatment, mean effects for time among GAD participants were significant for other psychological distress measures, but not for the BAI and the Hamilton Depression Rating Scale, suggesting that the BAI may not be sensitive to treatment change. Overall, the initial evidence of psychometric properties as well as the simplicity of the BAI makes it a useful tool to detect the presence of anxiety in older adults. However, because of (a) potential confounds with depressive symptoms and (b) high somatic item content (13 of the 21 items are related to somatic symptoms), the BAI should be used with caution, especially with samples recruited from medical settings. Additionally, more evidence is required before it is deemed acceptable for use in evaluating treatment effects.

General Health Questionnaire (GHQ)

The GHQ (Goldberg, 1978) is a self-report questionnaire designed to evaluate the presence of minor, non-psychotic psychiatric disorders in community setting. The original questionnaire consists of 60 items, but shorter versions of 30, 28, 20, and 12 items have also been developed. The 28-item version is most commonly used in the general population but the shorter 12-item version may be more appropriate with older populations [2](Clarke & Clarkson, 2009). The GHQ incorporates four scales: somatic symptoms, anxiety and insomnia, social dysfunction and severe depression. The items are rated on a 4-point scale ("not at all", "no more than usual", "rather more than usual" and "much more than usual") and are summed to provide a score ranging from 0 to 84, with higher scores representing higher levels of distress. When using the full 60-item scale, a cutoff of 23/24 for the total scale is suggested. However, Goldberg and Hillier (1979) suggest using an alternative binary scoring method in which the two least symptomatic answers ("not at all" and "no more than usual") are given a score of 0 and the two most symptomatic answers ("rather more than usual" and "much more than usual") a score of 1. A total score of 4 or more on any subscale suggests caseness. There is no specific cutoff suggested, nor are norms available, for older adults. However, some evidence has shown that when using the binary scoring method with a population of older adults, the best cutoff score for each subscale is 3/4 (Pappassotiropoulos, Heun, & Maier, 1997).

The GHQ was developed to be used with adults and adolescents. Although it is one of the most commonly used measures to detect the presence of mental disorders, very little is known about its use with older adults. The internal consistency of the measure is high in samples of older community dwelling adults when using either the 12 or 28-item version (α = .75 to .90;

[2]Its psychometric properties have been found to be equal or better than the longer version when used with older adults (i.e. Salama-Younes, Montazeri, Isma & Roncin, 2009).

Cheung, 2002; Clarke et al., 2009; Malakouti, Fatollahi, Mirabzadeh, & Zandi, 2007; Costa et al., 2006; Boey & Chiu, 1998), in cognitively impaired older adults when using the 12 item version (α = .81, Costa et al., 2006) and in older medical outpatients when using either the 12 or 30-item scale (α = 0.82-0.92; Dale, Saevareid, & Soderhamn, 2009; Thygesen, Saevareid, Lindstrom, Nygaard, & Engedal, 2008). We were unable to locate any study reporting evidence of test-retest reliability for older adults. Few studies have examined the validity of the GHQ with older adults. There were no studies identified in our review in which the GHQ was used to assess the effects of mental health treatments for older adults. As a result there is no evidence with respect to the treatment sensitivity of the GHQ. Although most studies found that the GHQ could differentiate between older adults with and without mental disorder (Casta et al., 2006; Malakouti et al., 2007; Mowry & Burvill, 1990; Seva, Sarasola, Merino, & Magallon, 1991), one found that it does not differentiate mentally ill patients from those with somatic illness (Malakouti et al., 2007). In light of the limited psychometric information available on the GHQ when used with older adults, the GHQ should be used with considerable caution when assessing anxiety symptoms in older adults.

Brief Symptom Inventory (BSI)

The BSI (Derogatis & Spencer, 1982) is a 53-item self-report questionnaire designed to assess the psychological distress of medical and psychiatric patients. It is a brief form of the Symptom Checklist-90-R and covers nine symptom dimensions (somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, and psychotism) and three global indices of distress (Global Severity Index, Positive Symptom Distress Index, and Positive Symptom Total). The global indices measure the level of symptomatology, intensity of symptoms, and number of reported symptoms. The items are rated on a 5-point scale ranging from 0 (not at all) to 4 (extremely). The overall score of the BSI is referred to as the Global Severity Index (GSI). Scores on the GSI ranges from 0 to 72 and each

subscale has a score ranging between 0 and 24. The scores are interpreted by comparing them to age appropriate norms. The manual provides norms for adult non-patients, psychiatric outpatients, inpatients, and adolescent non-patients, but not for older adults. However, Hale and colleagues (1984) have provided age-relevant norms for community dwelling older adults.

Although the BSI was developed for adolescents and adults, some research has been done on its use with older adults. High internal consistency has been found in samples of cognitively impaired older adults (α = .82; Fisher, Segal, & Coolidge, 2003), older medical outpatients (α = .89; Petkus et al., 2010; Petkus, Gum, King-Kallimanis, & Wetherell, 2009), community dwelling older adults (α = .72-.79; Fisher et al., 2003; Pektus et al., 2009), and older adult psychiatric outpatient (α = .72-.90; Wetherell et al., 2010; Wetherell et al., 2007). We found no study reporting test-retest reliability data for the BSI in samples of older adults. There has been little effort to validate the BSI with older adults. One study found that the BSI did not distinguish between medically ill older patients with and without an anxiety disorder (Wetherell et al., 2007), whereas another found that it could discriminate between homebound older adult with and without anxiety disorders (Petkus et al., 2010). There were no studies identified in our review in which the BSI was used to assess the effects of mental health treatments for older adults. Somatic symptoms of the BSI are grouped in a somatization subscale and therefore provide some assurance scores on anxiety-related subscales are not inflated by the presence of symptoms better explained by a medical condition. Although the BSI is simple and covers a wide range of symptoms, the lack of supporting psychometric evidence for older adults, especially validity evidence, severely limits its usefulness with older adults.

Penn State Worry Questionnaire (PSWQ)

The PSWQ (Meyer et al., 1990) is a 16-item self-report questionnaire designed to evaluate pathological worry. The items are rated on a 5-point scale and are summed to provide a score

ranging from 16 to 80, with higher scores reflecting higher levels of worry. Eleven items are worded in the direction of pathological worry, whereas the remaining items are worded to indicate the absence of worry. Although there are no specific cutoff scores, the mean score for individuals with generalized anxiety disorder is between 60 and 68. There are no norms for older adults, but a cutoff score of 50 has been suggested for use with samples of older medical patients (Stanley et al., 2003).

The PSWQ was developed and normed for younger adults, but several studies have examined its psychometric properties in samples of older adults. High internal consistency has been found in samples of older adults diagnosed with generalized anxiety disorder (α = .81-.89; Beck, Stanley, & Zebb, 1995; Stanley et al, 2001; Wetherell et al., 2003), in home care residents (α = .79; Diefenbach et al., 2009), and in community dwelling older adults (α = .80-.91; Beck et al., 1995; Hunt, Wisocki, & Yanko, 2003; Senior et al., 2007). Moderate to high test-retest reliability has been found in samples of older adults (r = .54-.78; Hopko et al, 2003; Stanley et al., 2001). The PSWQ showed adequate concurrent validity by virtue of significant correlations with other self-report measures of anxiety (r = .29-.79: Andreescu et al., 2008; Diefenbach et al., 2009; Hopko et al., 2003; Kogan et al., 2000; Stanley et al., 2001; Wetherell et al., 2003). Correlations with self-report measures of depression were lower and showed some evidence of discriminant validity (r = .12-.51; Diefenbach et al., 2009; Hopko et al., 2003; Kogan et al., 2000; Senior et al., 2007; Stanley et al., 2001; Wetherell et al., 2003). There is concern that some older adults have difficulty completing and interpreting the content of the reversed items of the PSWQ (Stanley et al., 2003; Wetherell et al., 2003). In order to respond to this problem, Hopko and his colleagues (2003) eliminated 8 of the original items to create an abbreviated scale (PSWQ-A). This scale has shown good psychometric properties in samples of older adults (e.g., Crittendon & Hopko, 2006; Hopko et al., 2003; Nuevo, Mackintosh, Gatz, Montorie & Wetherell, 2007).

Two studies identified in our review used the PSWQ to evaluate the effects of CBT for treating late-life anxiety. In a study examining the efficacy of a CBT treatment relative to a minimal contact control group in older adults with GAD, an analysis of simple effects demonstrated significant improvements on the PSWQ as well as in other anxiety measures for the CBT group (Stanley et al., 2003). In the aforementioned Wetherell et al. (2003) study, the main effects for time were significant for the PSWQ as well as most of the other anxiety and depression measures. These two studies are supportive of the treatment sensitivity of the PSWQ, but additional studies assessing the effects of mental health treatments for late-life anxiety with the PSWQ are needed. Overall then, although the lack of norms for older adults is problematic, initial evidence suggests that the PSWQ and its abbreviated form may be useful in assessing worry in older adults.

Symptom Checklist 90-R (SCL-90-R)

The SCL-90-R (Derogatis, 1994) is a 90-item self-report questionnaire designed to evaluate a wide range of psychological problems and symptoms of psychopathology. It covers nine primary symptom dimensions (somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, and psychotism) and has three global indices of distress that give an overall sense of psychological distress. The items are rated on a 5-point scale. A score of 50 or lower on any scale is considered to be in the normal range and a score of 65 or above is considered to be a clinical case. The cutoff scores have not been validated for use with older adults. The manual provides norms for adult and adolescents non-patients, adult psychiatric outpatients, and adult psychiatric inpatients; no norms are provided for older adults. Furthermore, we were unable to find any studies examining its psychometric properties with older adults and the reviewed articles that used the SCL-90-R did not report any psychometric information. There were no studies identified in our review in which

the SCL-90-R was used to assess the effects of mental health treatments for older adults. Overall, the lack of evidence for the psychometric soundness of the SCL-90-R with older adults means that it is currently not an appropriate measure for use with this population.

Worry Scale (WS)

The WS (Wisocki et al., 1986) is a 35-item self-report questionnaire developed to measure worries in the areas of health (17 items), finances (5 items) and social conditions (13 items). The items are rated on a 5-point scale ranging from 0 (never) to 4 (much of the time) and are summed to provide a score ranging from 0 to 140, with higher scores reflecting higher levels of worry. No norms or cutoff scores are provided in the original study. However, a mean total score of 35.4 for individual with GAD and of 11 for non-anxious control have been reported in the literature (Stanley, et al., 1996). Community samples of active older adults have yielded total mean scores ranging from 10.4 to 17.4 and samples of homebound older adults have found mean scores ranging from 17.3 to 23.7 (Wisocki, 1994).

The WS has been created specifically for use with older adults. The WS total score is associated with excellent overall internal consistency in both GAD (α = 0.81-0.93, Stanley et al., 1996; Stanley et al., 2001) and non-anxious samples (α = 0.93-0.94; Hunt et al., 2003; Stanley et al., 1996) and good internal consistency for its subscales (α = 0.76 to 0.95; Hunt et al., 2003; Stanley et al., 1996). Adequate test-retest reliability has been found in samples of older adults for the total scale (r= 0.69-0.70, Stanley et al., 2001; Stanley et al., 1996) and the subscales (r = 0.58-0.80; Stanley et al., 2001; Stanley et al., 1996). Concurrent validity for the WS has been shown by significant correlations between the scale and measures of anxiety (r = 0.54-0.63, Stanley et al., 2001; Wisocki, 1988, Wisocki et al., 1986). However, the scale also showed high correlations with measures of depression (r = 0.50 to 0.78 Wisocki, 1988; Wisocki et al., 1986). As noted

previously, examining a CBT intervention for GAD, Stanley et al. (2003) found that improvements were noted for the CBT group on the WS and other anxiety measures.

The original scale has been revised and expanded to include 88 items in six dimensions (finances, health, social conditions, personal concerns, family concerns and world issues). Its psychometric properties are currently under investigation, with initial evidence indicating that they are similar to the original scale (Hunt et al., 2000; Watari & Brodbeck, 2000). Overall then, although more research is needed on discriminant validity and treatment sensitivity, initial evidence suggests that the WS may be useful in assessing worry in older adults.

Geriatric Anxiety Inventory (GAI)

The GAI (Pachana et al., 2007) is a 20-item self-report questionnaire designed to measure anxiety symptoms in older adults. The questionnaire uses an agree/disagree response choice format, with the number of "agree" responses added for the total score. The maximum score is 20, with higher scores indicating higher anxiety. According to the original study, the optimal cutoff score to identify GAD in older adults is 10/11, and 8/9 to identify other anxiety disorders (Pachana et al., 2007). Similar results have been found in other studies of psychogeriatric patients (Byrne, Pachana, Goncalves, Arnold, King, & Khoo, 2010; Pachana et al., 2007).

The GAI was developed and normed with samples of community-dwelling older adults and older adults receiving psychiatric services. Excellent internal consistency has been shown in the original study of community-dwelling older adults ($\alpha = 0.92$; Pachana et al., 2006) and older adults receiving psychiatric services ($\alpha = 0.93$; Pachana et al., 2006). High internal consistency has also been found in other samples of community-dwelling older adults ($\alpha = 0.90$-0.92; Andrew & Dulin, 2007; Pachana et al., 2007; Byrne et al., 2010), in psychogeriatric samples ($\alpha = 0.93$; Pachana et al. 2007), and in an older adult sample receiving home care ($\alpha = 0.93$; Diefenbach et al., 2009). Lower, but still good, internal consistency was found in older adults with mild

cognitive impairment (α = 0.76; Rozzini et al., 2009). Researchers have described the data obtained from the instrument as having sound test-retest reliability (Pachana et al., 2007; Pachana et al., 2006), but not report precise reliability information. One study conducted with older adults with mild cognitive impairment showed good test-retest reliability (*r* = 0.86; Rozzini et al., 2009). Moderate to strong correlations between the GAI and other anxiety measures show evidence of concurrent validity, with correlations ranging from 0.58 to 0.86 (Pachana et al., 2007; Byrne et al., 2010; Diefenbach et al., 2009;). However, relatively high correlations (*r* = 0.65-0.79) between the GAI and depression measures provide limited evidence of discriminant validity (Andrew et al., 2006; Diefenbach et al., 2009; Paukert, 2009). There were no studies identified in our review in which the GAI was used to assess the effects of mental health treatments for older adults. Taken together, although much more research evidence is needed, initial evidence suggests that the GAI is likely to be a useful tool for detecting anxiety in older adults.

Discussion

The aims of this review were to evaluate the mental health research literature on older adults in order to (1) identify the most commonly used anxiety measures and (2) determine how appropriate the measures are for clinical or research use with older adults by examining their psychometric properties. Results of our literature search indicate that, with more than 90 measures used to evaluate anxiety in older adults, no clear consensus exists amongst researchers on which anxiety measure is more appropriate to use when evaluating a geriatric population. Out of the most commonly used measures, only one (GMSE) was developed specifically for older adults. Therefore, it seems that, with the lack of anxiety measures created and validated for older adults, researchers and clinicians often need to rely on anxiety measures created for younger adults when assessing older populations. However, because many differences exist between younger and older adults, it is unlikely that a single measure can adequately assess anxiety across

the entire adult life span. For example, compared to younger adults, older adults report more somatic symptoms, which means that anxiety measures that were not specifically developed for older adults may not provide an accurate assessment of anxiety (Fuentes & Cox, 1997; Wetherell & Gatz, 2005). As a result, it is extremely important for clinicians and researchers to have access to valid and reliable measures with established age-appropriate norms.

Most of the assessment measures we reviewed lack sufficient evidence for their psychometric soundness when used with older adults. Several critical considerations limit their use with geriatric populations. First, although each reviewed measure showed adequate internal consistency ($\alpha \geq .70$), the existing data often come from a single published study and mostly from samples of older psychiatric outpatients. Both replication and extension of previous findings are necessary in order to determine if the measures are good general screening measures for anxiety in older adults or whether they may have relevance only for a specific subgroup of older adults. Second, only four measures (BAI, GMSE, PSWQ, STAI) showed evidence of adequate test-retest reliability in older adult samples. The lack of test-retest reliability needs to be taken into consideration if the measure is to be used numerous times, such as when evaluating treatment effects. The third limitation resides in the lack of evidence for both discriminant and concurrent validity in most measures. Three measures showed adequate concurrent validity with older adults (GMSE, BAI, PSWQ), four showed limited concurrent validity (GADS, HARS, HADS, STAI) whereas no information was found for the three other measures. Every measure, except the PSWQ, either showed limited or low discriminant validity. Also, very few instruments have adequate normative data for older adults, which severely limit their clinical value. Only the GMSE has been normed for older adults and provides cutoff scores validated for this population. Finally, few measures had evidence of treatment sensitivity, suggesting the urgent need for

studies assessing the sensitivity of commonly used anxiety instruments to the effects of mental health treatments for older adults.

Taken together, three of the most commonly used measures showed sufficient psychometric evidence to warrant their use in assessing anxiety in older adults. The BAI, a measure of general anxiety, and the PSWQ, a measure of worry, have both demonstrated high internal consistency in older psychiatric outpatients as well as in community dwelling older adults. Test-retest reliability of the two measures has been shown to be adequate. Convergent validity for these measures has been suggested by moderate correlations with other self-report measures of anxiety. The PSWQ showed good discriminant validity with low correlations with self-report measures of depression but the BAI show only limited discriminant validity. Although both measures were initially developed for younger populations, subsequent evidence suggests that they might be good choices when selecting an anxiety measure for a geriatric population. The GMSE, a semi-structured interview created for older adults that includes anxiety clusters, is the only measure that provides norms and cutoff scores validated for older adults. That, combined with its appropriate psychometric properties, suggests it is a good option for assessing anxiety in older adults. However, the GMSE has not been used in research published in the last decade, which may suggest that more recent measures may be more appropriate to measure anxiety in older adults.

Indeed there is a growing number of anxiety instruments designed for use with older adults. Both the WS and the GAI were specifically developed for older adults and showed sufficient psychometric evidence to warrant their use in assessing anxiety in older adults. In many instances, these instruments might be preferred over the more commonly used measures that were not developed for the geriatric population. Both measures demonstrated high internal consistency in older psychiatric outpatients as well as in community dwelling older adults. Test-retest

reliability of the two measures has been shown to be good. Convergent validity for these measures has been provided by significant correlations with other self-report measures of anxiety. But, because of the extent to which these measures have been found to correlate with measures of depression, further research focusing on discriminant validity is certainly required.

Coexistence of Somatic Symptoms

In this review, six of the most commonly used measures (BAI, BSI, GHQ, GADS, HARS, SCL-90-R) were weighted heavily with somatic symptoms, which makes it difficult to distinguish between anxiety symptoms and symptoms of other health problems (or even normal aging) among the geriatric population. This can be problematic as the experience of anxiety varies greatly in younger and older adults. Not only do older adults experience more somatic symptoms when anxious but they are also likely to have coexisting physical conditions that may produce anxiety-like symptoms. It is therefore critical for the measures to be able to distinguish between anxiety-like symptoms caused by a medical condition from symptoms caused by an anxiety disorder. Ideally, measures used with older adults would be used in a manner that considered differential diagnoses in order to distinguish somatic symptoms caused by anxiety an physical illness. By using a measure that includes many somatic symptoms, a high proportion of non-anxious older adults experiencing symptoms of a medical condition may fall within the range used to identify clinical anxiety in a younger population. This must be taken into consideration when using these measures with older adults, particularly with older adults in medical settings or with a medical condition.

Coexistence of Depression

Another issue that must be taken into consideration when evaluating the presence of anxiety in older adults is the frequent coexistence of anxiety and depression in later life. Studies suggest that as much as 38% to 46% of older adults meeting criteria for a mood disorder also

meet criteria for an anxiety disorder (Beekman et al., 1998; Flint, 1994; Lenze et al., 2000). As these comorbid conditions can increase the complexity of anxiety assessment and diagnosis, it is important that the measures differentiate anxious and depressive symptoms in older adults. In this review, none of the most commonly used measures showed adequate evidence of discriminant validity with respect to mood disorders, which is likely to lead to a high misclassification rate among older adults. Two measures (GADS, HADS) include both a depression and an anxiety subscale but research indicates that, for both measures, the anxiety and depression subscales are highly correlated and therefore might not distinguish between the disorders (Davies et al., 1993; Flint et al., 2002; Kolowski et al., 2008). Consequently, researchers and clinicians should be careful when using these measures until more solid evidence of discriminate validity is obtained. Obviously information beyond what is available from these instruments must be considered when making any diagnostic formulations.

Limitations

Findings from this review must be interpreted within the limitations of systematic reviews in general. A key issue in selecting articles for review was deciding which studies to include and which to exclude. In particular, we required that studies have samples in which all participants were at least 65 years of age. We excluded from consideration, therefore, a number of articles in which some the research sample was included participants in their late 50s or early 60s. Accordingly, this limited the number of studies we examined in detail. Setting the inclusion criteria to allow samples in which all participants were at least 55 years of age would certainly have increased the number of studies available for our review, but at the cost of including too broad a range of ages for our intended focus on older adults. The search was also limited to articles published in either French or English as those were the languages understood by the authors. Furthermore, we searched for studies in the most important databases for psychology

(PsycInfo) and medicine (PubMed), but other databases were not considered. It is conceivable that this might have resulted in an overly narrow review of the published literature on anxiety measures used with older adults. However, our search strategy did ensure that we considered the vast majority of journals that publish research on the health and/or mental health of older adults.

Conclusion

The present systematic review shows that the anxiety measures most commonly used with older adults are mostly measures developed for a younger population. Although there is empirical support for the use of some of these measures, the majority of measures lack sufficient evidence of their psychometric soundness when used with older adults. The STAI was found to be the most commonly used measure in the reviewed articles. However, an examination of its psychometric properties yielded mixed results, suggesting that it does not yet show sufficient supporting psychometric evidence and should, therefore, be used with caution when assessing older adults. The HADS was also frequently used, but the lack of psychometric evidence for this instrument suggests that it is not be a good option when evaluating older adults until more research examines its validity and reliability. The GMSE was created for older adults and shows preliminary evidence of psychometric soundness. However, this measure has not been used in recent studies and, therefore, might not be the most appropriate choice when assessing anxiety in older adults. Both the PSWQ and the BAI have shown good psychometric properties, suggesting they may be useful tools to measure the presence of anxiety in older adults. However, considering the limited research on the psychometric properties of the measures, clinicians and researchers must be cautious and carefully consider the strengths and weaknesses of each measure before deciding which one to use for a specific purpose. Although the evidence to date is somewhat limited, the measures specifically developed for older adults, such as the WS and the GAI, should be

seriously considered by clinicians and researchers when assessing anxiety in a geriatric population.

A major shortcoming evident in the reviewed measures is the inclusion of somatic symptoms of anxiety that often overlap with the symptoms of normal aging, comorbid conditions, and medication side effects. The use of measures heavily weighted for somatic symptoms should be avoided in medical settings or in samples presenting with a medical condition. Caution should also be used when assessing older adults who present possible depressive symptoms, as many of the anxiety measures are highly correlated with measures of depression. Additionally, most measures do not present age appropriate norms or clinically relevant cut-off scores, which greatly limit their use with older adults. For the evaluation of anxiety in older adults to be more evidence-based, there is a pressing need for the validation of measures that were created for young adults and that are used with older adults. Research is needed on the development and validation of anxiety measures created specifically for older adults that evaluate a wide range of anxiety symptoms and disorders.

References

Alwahhabi, F. (2003). Anxiety symptoms and generalized anxiety disorder in the elderly: A review *Harvard Review of Psychiatry, 11*, 180-193.

American Psychological Association Presidential Task Force on Evidence Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.

Ames, D., Flynn, E., Tuckwell, V., & Harrigan, S. (1994). Diagnosis of psychiatric disorder in elderly general and geriatric hospital patients: AGECAT and DSM-III-R compared. *International Journal of Geriatric Psychiatry, 9*, 627-633.

Andreescu, C., Belnap, B.H., Rollman, B.L. Houck, P., Ciliberti, C., Mazumador, S., et al. (2008). Generalized anxiety disorder severity scale validation in older adults. *The American Journal of Geriatric Psychiatry*, *16*, 813-818.

Andrew, D. H., & Dulin, P. L. (2007). The relationship between self-reported health and mental health problems among older adults in New Zealand: Experiential avoidance as a moderator. *Aging and Mental Health, 11*, 576-603.

Ayers, C. R., Sorrell, J. T., Thorp, S. R. & Wetherell, J.L. (2007). Evidence-based psychological treatments for late-life anxiety, *Psychology and Aging, 22*, 8-17

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: psychometric properties. *Journal of Consulting and Clinical Psychology, 56,* 893-897.

Beck, A. T., & Steer, R.A. (1990). *Manual for the Beck Anxiety Inventory*. San Antonio, TX: Psychological Corporation.

Beck, J. G., Stanley, M. A., & Zebb, B. J. (1995). Psychometric properties of the PSWQ in older adults. *Journal of Clinical Geropsychology, 1*, 33-42.

Beck, J. G., Stanley, M. A., & Zebb, B. J. (1996). Characteristics of generalized anxiety disorder in older adults: A descriptive study. *Behaviour Research and Therapy, 34*, 225-234.

Beck, J. G., Stanley, M. A., & Zebb, B. J. (1999). Effectiveness of the Hamilton Anxiety Rating Scale with older generalized anxiety disorder patients. *Journal of Clinical Gerpsychology, 5*, 281-290.

Beekman, A. T., Bremmer, M. A., Deeg, D. J., van Balkom, A. J., Smit, J. H, DeBeurs, E., et al. (1998). Anxiety disorders in later life: A report from the longitudinal aging study Amsterdam. *International Journal of Geriatric Psychiatry, 13*, 717-726.

Boey, K. W., & Chiu, H. F. (1998). Assessing psychological well-being of the old-old: A comparative study of GDS-15 and GHQ-12. *Clinical Gerontologist, 19 (1),* 65-75.

Bryant, C., Jackson, H., & Ames, D. (2009). Depression and anxiety in medically unwell older adults: Prevalence and short-term course. *International Psychogeriatrics, 21*, 754-763.

Byrne, G., Pachana, N., Goncalves, D., Arnold, E., King, R., & Khoo, S. (2010). Psychometric properties and health correlates of the Geriatric Anxiety Inventory in Australian community-residing older women. *Aging and Mental Health, 14*, 247-254.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*, 7-18.

Cheung, Y. B. (2002). A confirmatory factor analysis of the 12-item General Health Questionnaire among older people. *International Journal of Geriatric Psychiatry, 17*, 739-744.

Christensen, H., Jorm, A. F., Mackinnon, A. J., Korten, A. E., Jacomb, P. A, Hendersen, A. S., & Rodgers, B. (1999). Age differences in depression and anxiety symptoms: A structural equation modelling analysis of data from a general population sample. *Psychological Medicine, 29*, 325-339.

Clarke, D., & Clarkson, J. (2009). A preliminary investigation into motivational factors associated with older adults problem gambling. *International Journal of Mental Health and Gambling, 7*, 12-28.

Cohen, L. L., La Greca, A. M., Blount, R. L., Kazak, A. E., Holmbeck, G. N., & Lemanek, K. (2008). Introduction to special issue: Evidence-based assessment in pediatric psychology. *Journal of Pediatric Psychology, 33*, 911-915.

Copeland, J. R., Beekman, A. T., Dewey, M. E., Hooijer, C., Jordan, A., Lawlor, B. A., et al. (1999). Depression in Europe: Geographical distribution among older people. *British Journal of Psychiatry, 174*, 312-321.

Copeland, J. R., Dewey, M. E., & Griffiths-Jones, H. M. (1986). A computerized psychiatric diagnostic system and case nomenclature for elderly subjects: GMS and AGECAT. *Psychological Medicine, 16*, 89-99.

Copeland, J. R. M., Kelleher, M., J., Duckwort, G., & Smith A. (1976). Reliability of psychiatric assessment in older patients. *International Journal of Aging and Human Development, 7*, 313-322

Copeland, J. R., Kelleher, M. J., Kellett, J. M., Gourlay, A. J., Cowan, D. W., Barron, G., et al. (1975). Cross-national study of diagnosis of the mental disorders: A comparison of the diagnoses of elderly psychiatric patients admitted to mental hospitals serving Queen's county, New York, and former Borough of Camberwell, London. *British Journal of Psychiatry, 126*, 11-20.

Copeland, J. R. M., Kelleher, M. J., Kellett, J. M., Gourlay, A. J., Gurland, B. J., Fleiss, & Sharpe, L. (1976). A semi-structured clinical interview for the assessment of diagnosis and mental state in the elderly: The Geriatric Mental State Schedule I: Development and reliability. *Psychological Medicine, 6*, 439-449.

Copeland, J. R., Prince, M., Wilson, K. C. , Dewey, M. E., Payne, J., & Gurland, B (2002). The Geriatric Mental State Examination in the 21st century. *International Journal of Geriatric Psychiatry, 17*, 729-732.

Costa, E., Barreto, S., Uchoa, E., Firmo, J., Lima-Costa, M., & Prince. M. (2006). Is the GDS-30 better that the GHQ-12 for screening depression in elderly people in the community? The Bambui Health Aging Study (BHAS). *International Psychogeriatrics, 18*, 493-503.

Crittendon, J. & Hopko, D.R. (2006). Assessing worry in older and younger adults: Psychometric properties of an abbreviated Penn State Worry Questionnaire. *Journal of Anxiety Disorders, 20*, 1036-1054.

Cully, J.A., Graham, D.P., Stanley, M.A., Ferguson, C.J., Sharafkhaneh, A., Souchek, J., & Kunik, E. (2006). Quality of life in patients with chronic obstructive pulmonary disease and comorbid anxiety or depression. *Psychosomatics, 47*, 312-319.

Dale, B., Saevareid, H., & Soderhamn, O. (2009). Testing and using Goldberg's General Health Questionnaire: Mental health in relation to home nursing, home help, and family care among older, care-dependent individuals. *International Journal of Mental Health Nursing, 18*, 133-143.

Davies, K. N., Burn, W. K., Mckenzie, F. R., Brotherwell, J. A & Wattis, J. P. (1993). Evaluation of the Hospital Anxiety and Depression Scale as a screening instrument in geriatric medical inpatients. *International Journal of Geriatric Psychiatry, 8*, 165-169.

de Beurs, E., Beekman, A. T., van Balkom, A. J., Deeg, D. J., van Dyck, R., & van Tilburg, W. (1999). Consequences of anxiety in older persons: Its effect on disability, well-being and use of health services. *Psychological Medicine, 29*, 583-593.

Dennis, R. E., Boddington, S. J., & Funnell, N. J. (2007). Self-report measures of anxiety: Are they suitable for older adults? *Aging & Mental Health*, 11, 668-677.

Derogatis, L. R. (1994). *Administration, scoring and procedures manual*. Minneapolis, MN: National Computer System.

Derogatis, L. R., & Spencer, P. M. (1982). *The Brief Symptom Inventory (BSI): Administration, and procedures manual-I*. Baltimore, MD: Clinical Psychometric Research.

Diefenbach, G. J., Stanley, M. A., Beck, J. G., Novy, D. M., Averill, P. M., & Swann, A.C. (2001). Examination of the Hamilton scales in assessment of anxious older adults: A replication and extension. *Journal of Psychopathology and Behavioral Assessment, 23*, 117-124.

Dienfenbach, G., Tolin, D., Meunier, S., & Gilliam, C. (2009). Assessment of anxiety in older home care recipients. *The Gerontologist, 49*, 141-153.

Edelstein, B. A., Woodhead, E. L., Segal, D. L., Heisel, M. J., Bower, E. H., Lowery, A. J., & Stoner, S.A. (2008). Older adult psychological assessment: Current instrument status and related considerations. *Clinical Gerontologist, 31(3)*, 1-35.

Fisher, B. M., Segal, D. L., & Coolidge, F. L. (2003). Assessment of coping in cognitively impaired older adults: a preliminary study. *Clinical Gerontologist 26(3)*: 3-12.

Flint, A. J.(1994). Epidemiology and comorbidity of anxiety disorders in the elderly. *The American Journal of Psychiatry, 151,* 640-649.

Flint, A. J., & Rifat, S. L. (2002). Factor structure of the Hospital Anxiety and Depression Scale in older patients with major depression. *International Journal of Geriatric Psychiatry, 17*, 117-123.

Fuentes, K. & Cox, B. J. (1997). Prevalence of anxiety disorders in elderly adults: A critical analysis. *Journal of Behavior Therapy and Experimental Psychiatry, 28,* 269-279

Forman, S., Fagley, N. S., Steiner, D. D., & Schneider, K. (2009). Teaching evidence-based interventions: Perceptions of influences on use in professional practice in school psychology. *Training and Education in Professional Psychology, 3*, 226-232.

Goldberg D. et al. (1978). *Manual of the General Health Questionnaire*. Windsor, England: NFER Publishing.

Goldberg, D., Bridges, K., Duncan-Jones, P., & Grayson, D. (1988). Detecting anxiety and depression in general medical settings. *British Medical Journal, 297*, 897-899.

Goldberg, D. P., & Hillier, V. F. (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine, 9*, 139-145.

Hale, D. W., Cochran C. D., & Hedgepeth, B. E. (1984). Norms for the elderly on the Brief Symptom Inventory. *Journal of Consulting and Clinical Psychology, 52*, 321-322.

Hamilton, M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology, 32*, 50-55.

Haworth, J. E., Moniz-Cook, E., Clark, A. L, Wang, M., & Cleland, J. G. (2007). An evaluation of two self-report screening measures for mood in and out-patient chronic heart failure population. *International Journal of Geriatric Psychiatry*, *22*, 1147-1153.

Henderson, A. S., Duncan-Jones, P., & Finlay-Jones, R. A. (1983). The reliability of the Geriatric Mental State Examination. Community survey version. *Acta Psychiatrica Scandinavica, 67*, 281-289.

Himmelfarb, S., & Murrell, S. A. (1984). The prevalence and correlates of anxiety disorders in older adults. *Journal of Psychology, 116*, 159-167.

Holmbeck G. N., & Devine, K. A. (2009). Editorial: An author's checklist for measure development and validation manuscripts. *Journal of Pediatric Psychology, 34,* 691-696.

Hopko, D. R., Stanley, M. A., Reas, D. L., Wetherell, J. L., Beck, J. G., Novy, D.M., & Averill, P.M. (2003). Assessing worry in older adults: Confirmatory factor analysis of the Penn State Worry Questionnaire and psychometric properties of an abbreviated model. *Psychological Assessment, 15*, 173-183.

Huber, P., Mulligan, R., Mackinnon, A., Nebuloni-French, T., & Michel, J.P. (1999). Detecting anxiety and depression in hospitalised elderly patients using a brief inventory. *European Psychiatry, 14*, 11-16.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3,* 29-51.

Hunsley, J. & Mash E. J. (2008). *A guide to assessment that work.* New York: Oxford University Press.

Hunt, S., Wisocki, P., & Yanko, J. (2003). Worry and use of coping strategies among older and younger adults. *Journal of Anxiety Disorders, 17*, 547-560.

Institute of Medicine (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Johnson, G., Burvill, P. W., Anderson, C. S., Jamrozik, K. K., Stewart-Wynne, E. G., & Chakera, T. M. (1995). Screening instruments for depression and anxiety following stroke: Experience in the Perth community stroke study. *Acta Psychiatrica Scandinavica, 91*, 252-257.

Johnston, M., Pollard, B., & Hennessey, P. (2000). Construct validation of the hospital anxiety and depression scale with clinical populations. *Journal of Psychosomatics Research, 48*, 579-584.

Kabacoff, R. L., Segal, D. L., Hersen, M., & Van Hasselt V. B. (1997). Psychometric properties and diagnostic utility of the Beck Anxiety Inventory and the State-Trait Anxiety Inventory with older adult psychiatric outpatients. *Journal of Anxiety Disorder, 11,* 33-47.

Kenn, C., Wood, H., Kucyj, M., Wattis, J., & Cunane, J. (1987). Validation of the Hospital Anxiety and Depression Rating Scale (HADS) in an elderly psychiatric population. *International Journal of Geriatric Psychiatry, 2*, 189-193.

Kogan, J. N., Edelstein, B. A., & McKee, D. R. (2000). Assessment of anxiety in older adults: Current status. *Journal of Anxiety Disorders, 14*, 109-132.

Koloski, N. A., Smith, N., Pachana, N. A & Dobson, A. (2008). Performance of the Goldberg Anxiety and Depression Scale in older women. *Age and Ageing, 37*, 464-467.

Kvall, K., Macijauskiene, J., Engedal, K., & Laake, K. (2001). High prevalence of anxiety symptoms in hospitalized geriatric patients. *International Journal of Geriatric Psychiatry, 16*, 690-693.

Kvall, K., Ulstein, I., Nordhus, I., & Engedal, K. (2005). The Spielberger State-Trait Anxiety Inventory (STAI): The state scale in detecting mental disorders in geriatric patients. *International Journal of Geriatric Psychiatry, 20*, 629-634.

Lenze, E. J., Mulsant, B. H., Shear, M. K., Shulberg, H. C., Dew, M. A., Begley, et al. (2000). Comorbid anxiety disorders in depressed elderly patients. *The American Journal of Psychiatry, 157*, 722-728.

Lenze, E. J., Rollman, B. L., Shear, M. K., Dew, M. A., Pollock, B. G., Ciliberti, et al. (2009). Escitalopram for older adults with generalized anxiety disorder: A randomized controlled trial. *Journal of the American Medical Association, 301*, 295-303.

Malakouti, S. K, Fatollahi, P., Mirabzadeh, A., & Zandi, T. (2007). Reliability, validity and factor structure of the GHQ-28 used among elderly Iranians. *International Psychogeriatrics, 19*, 623-634.

Mendlowicz, M.V., & Stein, M.B. (2000). Quality of life in individuals with anxiety disorders. *American Journal of Psychiatry, 31*, 1602-1607.

Meyer, T., Miller, M., Metzger, R., & Borkovec, T.D. (1990). Development and validition of the Penn State Worry Scale. *Behaviour Research and Therapy, 28*, 487-495.

Morin, C. M., Landreville, P., Colecchi, C., McDonald, K., Stone J., & Ling, W. (1999). The Beck Anxiety Inventory: Psychometric properties with older adults. *Journal of Clinical Geropsychology, 5*, 19-29.

Mottram., P., Wilson, K., & Copeland, J. (2000). Validation of the Hamilton Depression Rating Scale and Montgomery and Asberg Rating Scales in terms of AGECAT depression cases. *International Journal of Geriatric Psychiatry, 15*, 1113-1119.

Mowry, B. J., & Burvill, P. W. (1990). Screening the elderly in the community for psychiatric disorder. *Australian and New Zealand Journal of Psychiatry, 24*, 203-206.

Nuevo, R., Mackintosh, M., Gatz, M., Montorio, I., & Wetherell, J.L. (2007). A test of the measurement invariance of a brief version of the Penn State Worry Questionnaire between American and Spanish older adults. *International Psychogeriatrics, 19*, 89-101.

Pachana, N.A., Byrne, G.J., Siddle, H., Koloski, N., Harley. E., & Arnold, E. (2007). Development and validation of the Geriatric Anxiety Inventory. *International Psychogeriatrics*, 19, 103-114.

Papassotiropoulos, A., Heun, R., & Maier, W. (1997). Age and cognitive impairment influence the performance of the general health questionnaire. *Comprehensive Psychiatry, 38*, 335-340.

Paukert, A. (2009). The roles of social support, self-efficacy, and optimism in physical health's impact on depressive and anxious symptoms among older adults. *Dissertation Abstracts International: Section B: The Science sand Engineering, 69*, 5788.

Petkus, A. J., Gum, A. M., King-Kallimanis, B., & Wetherell, J. L. (2009). Trauma history is associated with psychological distress and somatic symptoms in homebound older adults. *The American Journal of Geriatric Psychiatry*, *17*, 810-818.

Petkus, A. J., Gum, A. M., Small, B., Malcarne, V. L., Stein, M. B. & Wetherell, J. L. (2010). Evaluation of the factor structure and psychometric properties of the Brief Symptom Inventory-18 with homebound older adults. *International Journal of Geriatric Psychiatry, 25*, 578-587.

Pinquart, M., & Duberstein, P. R. (2007). Treatment of anxiety disorders in older adults: a meta-analytic comparison of behavioural and pharmacological interventions. *American Journal of Geriatric Psychiatry, 15,* 639-651.

Rozzini, L., Chilovi, B.V., Peli, M., Conti, M., Rozzini, R., Trabucchi, M., & Padovani, A. (2009). Anxiety Symptoms in mild cognitive impairment. *International Journal of Geriatric Psychiatry, 24*, 300-305.

Salama-Younes, M., Montazeri, A., Isma, A., & Roncin, C. (2009). Factor structure and internal consistency of the 12-item general health questionnaire (GHQ-12) and the Subjective Vitality Scale (VS), and the relationship between them : A study from France. *Health and Quality of Life Outcomes, 7*, 1-22.

Scheick, J. I. (1991). *Anxiety in the elderly: Treatment and research.* New York: Spring Publishing Co.

Schuurmans, J., Comijs, H., Emmelkamp, P. M., Weijnen, I. J., van den Hout, M, & Van Dyck, R. (2009). Long-term effectiveness and prediction of treatment outcome in cognitive

behavioral therapy and sertraline for late-life anxiety disorders. *International Psychogeriatrics, 21, 1148-1159.*

Segal, D. L., June, A., Payne, M., Coolidge, F. L., & Yochim, B. (2010). Development and initial validation of a self-report assessment tool for anxiety among older adults: The Geriatric Anxiety Scale. *Journal of Anxiety Disorders*, 24, 709-714.

Senior, A., Kunik, M., Rhoades, H., Novy, D., Wilson, N., & Stanley, M. (2007). Utility of telephone assessments in an older adult population. *Psychology and Aging, 22*, 392-397.

Seva, A., Sarasola, A., Merino, J.A., & Magallon, R. (1991). Validity of the scaled version of the General Health Questionnaire in a geriatric Spanish population. *The European Journal of Psychiatry, 5*, 32-36.

Skopp, N. A., Novy, D., Kunik, M., Daza, P., Adams, J. H., Senior, A., & Stanley, M. (2006). Investigation of cognitive behavior therapy. *The American Journal of Geriatric Psychiatry, 14*, 292.

Snaith, R. P., & Zigmond, A. S. (1994). *The Hospital Anxiety and Depression Scale Manual*. NFER, Nelson, Windsor.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P., & Jacobs, G. A. (1983). *Manual for the state-trait anxiety inventory STAI (Form Y): Self-evaluation questionnaire*. Palo Alto, CA: Consulting Psychologists Press.

Spinhoven, P. H., Ormel, J., Sloekers, P. P., Kempen, G. I., Speckens, A. E., & Van Hemert, A. M. (1997). A validation study of the hospital anxiety and depression scale (HADS) in different groups of Dutch subjects. *Psychological Medicine, 27,* 363-370.

Stanley, M.A., Beck, J. G., Novy, D. M., Averill, P. M., Swann, A. C., Diefenbach, G.J., & Hopko, D.R. (2003). Cognitive-behavioral treatment of late-life generalized anxiety disorder. *Journal of Consulting and Clinical Psychology, 71*, 309-319.

Stanley, M. A., Beck, J. G., & Zebb, B. J. (1996). Psychometric properties of four anxiety measures in older adults. *Behaviour and Research Therapy, 34,* 827-838.

Stanley, M. A., Novy, D. M., Bourland, S. L., Beck, J. G., & Averill, P. M. (2001). Assessing older adults with generalized anxiety: A replication and extension. *Behaviour Research and Therapy, 39,* 221-235.

Stanley, M. A., Wilson, N. L., Novy, D. M., Rhoades, H. M., Wagener, P. D., Greisinger, A.J., et al. (2009). Cognitive-behavior therapy for generalized anxiety disorder among older adults in primary care: A randomized trial. *Journal of the American Medical Association, 301*, 1460-1467.

Turina, C., Perdona, G., Bianchi, L., Cordioli, L., Burti, L., Micciolo, R., & Copeland, J. R. (1991). Inter-observer reliability of the Italian version of the Geriatric Mental State Examination. *International Journal of Geriatric Psychiatry, 6*, 647-680.

Thygesen, E., Saevareid, H. J., Lindstrom, T. C., Nygaard, H. A., & Engedal, K. (2009). Predicting needs for the nursing home admission – does sense of coherence delay nursing home admission in care dependant older people? A longitudinal study. *International Journal of Older People Nursing, 4*, 12-21.

Watari, K. F., & Brodbeck, C. (2000). Culture, health, and financial appraisals: Comparison of worry in older Japanese Americans and European Americans. *Journal of Clinical Geropsychology, 6*, 25-39.

Wetherell, J. L., & Areàn, P. A. (1997). Psychometric evaluation of the Beck Anxiety Inventory with older medical patients. *Psychological Assessment, 9,* 136-144.

Wetherell, J.L., Gatz, M. & Craske, M.G. (2003). Treatment of generalized anxiety disorder in older adults. *Journal of Consulting and Clinical Psychology, 7,* 31-40.

Wetherell, J. L., & Gatz, M. (2005). The Beck Anxiety Inventory in older adults with generalized anxiety disorder. *Journal of Psychopathology and Behavioral Assessment, 27*, 17-24.

Wetherell, J. L., Birchler, G. D., Ramsdell, J., & Unutzer, J. (2007). Screening for generalized anxiety disorder in geriatric primary care patients. *International Journal of Geriatric Psychiatry, 22*, 115-123.

Wetherell, J. L., Ayers, C. R., Nuevo, R., Stein, M. B., Ramsdell J. & Patterson, T. L (2010). Medical conditions and depressive, anxiety and somatic symptoms in older adults with and without generalized anxiety disorder. *Aging and Mental Health, 14*, 764-768.

Wisocki, P. A. (1988). Worry as a phenomenon relevant to the elderly. *Behavior Therapy, 19*, 369-379.

Wisocki, P. A., Handen, B., & Morse, C. K. (1986). The Worry Scale as a measure of anxiety among homebound and community active elderly. *The Behavior Therapist, 9,* 91-95.

Wisocki, P. A. (1994). The experience of worry among the elderly. In Davey, G.C.L. & Tallis, F. (Eds), *Worrying : Perspectives on theory, assessment and treatment* (pp. 246-261). New York: Wiley.

Wolitsky-Taylor, K., Castriotta, N., Lenze E. J., Stanley, M. A., & Craske, M. G. (2010). Anxiety disorders in older adults: A comprehensive review. *Depression and Anxiety, 27*, 190-211.

Yu, D. S., Lee, D.T., Woo, J., & Hui, E. (2007). Non-pharmacological interventions in older people with heart failure: Effects of exercise training and relaxation therapy. *Gerontology, 53*, 74-81.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression scale. *Acta Psychiatrica Scandinavica, 67*, 361-370.

Assessment of Anxiety in Older Adults: A Reliability Generalization Meta-Analysis of Commonly Used Measures*

Zoé Therrien* John Hunsley

University of Ottawa, Canada

* This study is in press in *Clinical Gerontologist*

Abstract

We conducted a reliability generalization meta-analysis of the 12 most commonly used measures of anxiety in older adults aged 65 and above. Of the 136 articles considered for inclusion, only 24% of published studies reported reliability coefficients from their original data collection. Sixty-three reliability coefficients from fifty-one articles and 16,183 individuals provided internal consistency reliability estimates for this meta-analysis. We present the average score reliabilities for each of the 12 measures, characterize the variance in score reliabilities across studies and consider sample and study characteristics that are predictive of score reliability. We discuss the importance of considering factors specific to the assessment of older adults (e.g., the frequency of a comorbid medical condition) as well as the importance of conducting sample specific reliability analyses. Recommendations are provided for researchers and clinicians choosing a measure of anxiety for use with older adults.

Keywords: anxiety, older adults, reliability, meta-analysis. Reliability generalization

Assessment of Anxiety in Older Adults: A Reliability Generalization Meta-Analysis of Commonly Used Measures

Anxiety disorders are highly prevalent, with rates typically exceeding those of depression and other affective disorders. Two to twelve months prevalence estimates of late life anxiety ranges from 1 to 15% in community samples of older adults (Bryant, Jackson, & Ames, 2009; Wolitsky-Taylor, Castriotta, Lenze, Stanley, & Craske, 2010) and from 1 to 28% in clinical samples of older adults (Bryant et al., 2009). Furthermore, among hospitalized geriatric patients, the prevalence in the preceding 3 days is estimated to be as high as 43% and is known to remain high after discharge (Kvall, Macijauskiene, Engedal, & Laake, 2001). Studies suggest that late-life subclinical anxiety is even more common, with 2-month prevalence estimates ranging from 15 to 52% in community samples and 15 to 56% in clinical samples (Bryant et al., 2009). This is particularly concerning, as older adults who present some anxiety symptoms are as negatively affected in their quality of life as those who meet criteria for an anxiety disorder (Preisig, Merikangas, & Angst, 2001; van Zelst, De Beurs, Beekman, Van Dyck, & Deeg 2006). The negative personal and societal costs associated with late-life anxiety disorders, including increased mortality, increased health care use, poor health, and decreased life satisfaction, coupled with the phenomenon of population aging, makes further research in this area a priority (Segal, June, Payne, Coolidge, & Yochim, 2010).

In both clinical and research contexts, anxiety appears to be measured primarily with self-report instruments (Alwahhabi, 2003; Dennis, Boddington, & Funnell, 2007). Numerous measures have been developed to gather information on the experience of anxiety and it can be a challenging task for researchers and clinicians to choose between them when looking for an anxiety measure. Additionally, as most of the assessment measures have not been created for

older adults, it is often necessary to use measures developed for younger populations of adults when working with older adults. However, when working with older adults, it is important to keep in mind that factors such as changes due to aging, the presence of physical symptoms and frequent comorbid disorders could significantly affect the validity and utility of self-report measures developed for younger adults. As a result, it is important to consider that many factors may complicate the use of self-report measures and that some existing measures may not be appropriately suited for research with older adults.

In a recent review of the research literature, Therrien and Hunsley (2012) found that there were a dozen anxiety measures commonly used in the assessment of older adults. See Table 1 for a list of the measures. Therrien and Hunsley concluded that (a) the anxiety measures most commonly used with older adults are mostly measures developed for a younger population and (b) although there is empirical support for the use of some of these measures, the majority lacked sufficient evidence of their psychometric soundness when used with older adults. In order to address the question of the extent of psychometric evidence for these measures, in the present study we examine the reliability values produced by these twelve commonly used measures of anxiety in a reliability generalization (RG) meta-analysis. We use meta-analytic data to test whether measures have the tendency to produce reliable scores across various sample characteristics. The results of this meta-analysis will provide valuable information for researchers seeking to identify an appropriate measure of anxiety for their own research with geriatric populations and for clinicians looking to select anxiety measures for assessing their older clients.

Reliability and Reliability Generalization

Reliability can be defined as the "degree to which test scores are free from errors of measurement" (American Psychological Association, 1985, p.19). According to classical psychometric theory, part of the variance of a set of scores is explained by the characteristics

measured in the test (true score) but other influences, like sampling or measurement error, can also explain a part of this variance (error score) (Graham, Yenling, & Jeziorski, 2006). Reliability estimates demonstrate the proportion of variance in score that is explained by the true score itself and therefore express the degree of consistency in the measurement of test scores. The reliability of generated scores is critical to the determination of a measure's merit as a measure cannot be scientifically sound without evidence of reliability.

Many researchers and clinicians erroneously believe that reliability is a property of a measure and that, once found reliable or unreliable, this status is unchangeable. However, a measure in itself is not reliable or unreliable; rather, reliability is a property of the scores obtained by a certain sample on the measure (Henson, Kogan, & Vacha-Haase 2001; Hunsley & Mash, 2008; Rousse, 2007; Vacha-Haase, Henson, & Caruso, 2002). In other words, a measure can produce reliable scores in one study and unreliable scores in another study with different participants. Reliability is affected not only by the quality of the measure itself but also by the composition and variability of the sample (e.g., age, gender, socio-economic status), sample size, administration and scoring of the measure (Barnes, Harp, & Jung, 2002).

Given the sample-specific nature of reliability, it is essential for researchers to calculate and report reliability coefficients from their own set of data, as failing to do so can lead to the interpretation of unreliable data (Henson et al., 2001). However, it is commonly found that not only do researchers often fail to report reliability coefficients for their own data but that they rely on previous findings of acceptable reliability derived either from previous studies or the original test manual and imply that their data are equally reliable (Twiss, Seaver, & McCaffrey, 2006; Kvaal, McDougall, Brayne, Matthews, & Dewey, 2008). This misleading assumption that the scores of a measure will necessarily be reliable because of scores obtained in previous sample has been termed *reliability induction* (Vacha-Haase, Kogan, & Thompson, 2000). Before assuming

the generalizability of reliability finds, users of a measure need to determine whether the reliability estimates reported in previous samples have comparable sample composition and score variability to the group for which the measure will be used (Kieffer & Reese, 2002). Even in cases where sample composition and score variability appear similar, score reliability incongruence is still possible.

Because of the sample-specific nature of reliability, Vacha-Haase (1998) proposed a meta-analytic method known as reliability generalization (RG) in order to estimate the mean reliability score of a specific measure based on the obtained reliability coefficients of different studies. This technique has been shown to be a powerful tool, and an increasing number of studies have used an RG meta-analysis in the last few years to examine the reliability of instruments designed to assess a variety of psychological construct (e.g., Lopez-Pina, Sanchez-Meca &Rosa-Alcazar, 2009; Vassar & Bradley, 2010). RG methodology involves the collection of all empirical research using the measure in question. When all articles have been retrieved, they are examined in order to identify studies in which researchers calculated reliability estimates for their own data. The reliability coefficients for these studies are recorded and various sample and study characteristics (e.g., sample size, age of the sample, percentage of males in the sample) are also coded.

RG meta-analyses have several important roles. In the first place, they provide the mean score reliability produced by a specific measure across studies. As the reliability estimate is not based on a single observation but rather on a large number of studies, it provides an important guideline for clinicians and researchers to consider when selecting a measure for a specific assessment task. RG meta-analyses also evaluate the variability in score various reliability produced by the measure across studies. By doing so, clinicians and researchers also have the confidence intervals around the mean reliability of scores produced by the measure which

provides stronger information than a single estimate. Finally, when significant variability exist between the reliability of scores produced by a measure, the RG can determine which variables can predict or explain the variations in score reliability for a specific measure. Thus RG meta-analysis allows researchers and clinicians to assess how the reliability of the scores produced by the measure can vary across sample and study characteristics (e.g. sample size, mean age of the sample, percentage of males in the sample).

Given that many of the anxiety measures used with older adults were not created specifically for an older population and that previous studies have questioned their psychometric properties (e.g., Dennis et al., 2007), the present study uses a RG method to estimate the mean reliability score of commonly used measures based on the obtained reliability coefficients of published studies. The RG will serve several important roles in assessing whether these anxiety measures produce reliable scores on responses provided by older adults. By reviewing the obtained score reliabilities of different studies for a specific measure, the RG analyses will (a) provide the mean score reliability produced by a each measure when used with an older population and (b) evaluate the variability in score reliability across samples for the measure. If there is substantial variance in the reliability of scores produced by a measure, then RG will determine which sample or study characteristics predicts or explains the variation in score reliability.

Method

Sample of Instruments and Articles

To prepare for the RG analyses, we used the databases PsycINFO and PubMed to identify the anxiety measures most commonly used with older adults in the literature. Studies meeting the following criteria were selected: (a) an empirical study, (b) used an anxiety measure with at least one participant, (c) only included adults aged 65 years and above in the sample and (d) written in

either English or French (the two languages understood by the authors). Empirical studies can be defined as research articles that report the results of a study that uses data derived from observation or experimentation. Additionally, the age cutoff follows the precedent set by previous researchers who set a cutoff of 65 years and up (e.g., Kvaal et al, 2008; Brenes et al., 2007) to define older adults. The search included articles published between the years 1960 and August 2009. To ensure the broadest possible search of the databases, two separate searches combining different key words and study criteria were carried out using the same keyword in both databases. The first search included the following terms: *anxiety*, *anxiety disorder, generalized anxiety disorder* AND *assessment*, *measurement*. The second search combined the following terms *anxiety*, *anxiety disorders*, *generalized anxiety disorder* AND *geriatric assessment*, *geriatric patients*, *geriatric psychiatry*, *geriatric*, *gerontology*.

After discarding duplicates, we identified a total of 1,427 articles (785 in PubMed and 642 in PsycINFO) for potential inclusion. After reviewing the titles and abstracts, 592 unique articles were retained for further examination. For all of the 592 selected articles, the full article was obtained and reviewed to assess the fit with our inclusion criteria. After this detailed review, a total of 213 articles were retained. Articles were excluded because they did not meet our inclusion criteria and more precisely did not have a sample that included only older adults (50%), did not use an anxiety measure (9%), did not have a sample of older adults and did not have an anxiety measure (7%), were not empirical studies (26%) and were not in French or English (5%), Other studies were unavailable in print or electronic form from within our university network (3%).

The details of the 213 retained articles were coded to summarize the anxiety measures used in the study, the reported reliability coefficient as well as key aspects of the study. A total of 91 different anxiety measures were used in the 213 articles. However, the majority of these

measures were used in only one or two studies. We focused on measures that had been frequently used in the literature, thus allowing independent replication of findings and the availability of data from multiple types of samples. With this in mind, and in consideration of the measure frequency distribution we found in the 213 studies, we decided to concentrate on instruments that had been used in six or more studies. As many studies did not report the reliability estimates for their own set of data, using this criterion ensured that all measures with sufficient reported reliability estimates were included in the RG analysis while infrequently used measures without sufficient information regarding reliability were excluded. However, as there has been considerable effort in the past decade to develop anxiety measures specifically suited for older adults, such measures are relatively new and might not yet be frequently used in the literature. In order to include all possible measures in the RG, we relaxed our criteria to consider measures developed in the past decade that had been used less than 6 times but that presented with sufficient frequency to allow for some replication of psychometric findings. Other measures could not be included as they did not present sufficient information regarding reliability to warrant their use in the RG. For example, the Brief Measure of Worry Severity (BMWS: Gladstone et al., 2005) was recently developed but was only used once in the literature, making it impossible to include in the analysis. Using this criterion, twelve measures were identified as being commonly employed in the literature (see Therrien & Hunsley, 2012). Only the articles that included at least one of these twelve measures were kept for the RG analyses, yielding a final sample size of 136 studies.

Of the 136 articles considered for inclusion, 51% made no mention of the reliability of their scores, 18% induced that a measure was reliable without reporting a value and 7% induced that a measure was reliable and provided a value from a different study. Only 24% of published studies reported reliability coefficients from their original data collection. This is substantially

lower than what was reported by Vacha-Haase, Ness, Nilsson, and Reetz (1999) who reviewed three journals and found that only 36% of the quantitative articles provided reliability coefficients for the data being analyzed. Recent RG analyses have found similar rates to that reported by Vacha-Haase et al. (1999) (i.e.Graham, Diebels, & Barlow, 2011; Lopez-Pina, Sanchez-Meca & Rosa-Alcazar, 2009; Wheeler, Vassar, Worley, & Barnes, 2011). In order to ensure that the reliability estimates were consistent and that the measurement of reliability was conducted similarly across studies, we focused on using Cronbach's alpha data. However, as the GMSE is a semi-structured clinical interview, we used inter-rater consistency as measured by kappa for this measure. Results obtained for the GMSE should therefore not be compared to the other measures as they do not represent the same reliability coefficient. In order to avoid dependence on the data used in the RG analyses, we ensured that every reliability coefficient came from a different sample. Thus, for example, if a study had a pretest and a post-test, only the reliability coefficients obtained at the pretest was included in our analysis.

Estimating Reliability

A difficulty in conducting an RG analysis resides in the fact that many researchers fail to report reliability estimates for their data set (Vacha-Haase et al., 2002; Vassar & Crosby, 2008). To compensate for this, Lane, White & Henson (2002) extended the typical methodology of RG by using the KR-21 technique to estimate the reliability of a set of data when the authors did not report it. The KR-21 technique, developed by Kuder and Richardson (1937), is an accepted general measure of internal consistency that requires the knowledge of the mean, the standard deviation and the number of items in a measure (Henson et al., 2001). The technique provides a value equivalent to coefficient alpha when a measure is made of dichotomous items scored 0 or 1 (Charter, 2007). We used the KR-21 technique to estimate the coefficient alpha of the GAI and the GADS, as they both use a dichotomous scale. This gave us an additional five reliability

estimates. In order to ensure that the values obtained by the KR-21technique were equivalent to the coefficient alpha obtained in other studies we conducted separate analysis of the two types of data (coefficient alphas and values obtained by KR-21) and statistically compared the obtained results. As no significant difference was found between the two reliability estimates, we combined all coefficients regardless of their type to pursue the analysis.

We also contacted the corresponding authors of every article in which the reliability for their set of data was not reported. Authors were asked to either send us information on the reliability coefficient for their data or send their dataset so that we could calculate the reliability coefficient. E-mails were sent to a total of 125 authors and we received a total of 26 reliability coefficients (a response rate of 21%). A total of 33 reliability estimates were obtained from the original studies, 5 reliability estimates were calculated from the KR-21 technique and 26 reliability estimates were obtained directly from the authors for a total of 64 reliability estimates included in the RG. Accordingly, of the 136 articles considered for inclusion, we obtained reliability coefficients for 53 articles representing 57 samples and 16,144 individuals and yielding 64 reliability coefficients[3]. It is important to note that the number of articles included in the RG and the number of reliability of coefficients are not the same, as some studies reported a reliability estimate for more than one anxiety measure. Additionally, as some studies presented more than one sample (e.g., community dwelling and medical outpatients), the number of reliability coefficient is not equal to the number of articles.

[3] Some of the measures included in this study do not consist of anxiety-specific measures, but rather measures of general distress or psychopathology that include anxiety subscales (i.e. BSI, GADS, GHQ,HADS, SCL-90). For these measures, only the reliability coefficients for the anxiety subscales were coded.

Coding of Study Characteristics

To examine potential relationships between the reliability estimates and the study features, both the Cronbach's alpha and possible moderator variables related to the instrument and the study participants were coded. According to psychometric theory, it can be expected that variables such as test length and standard deviation of the test scores will have an impact on the reliability of the measure, so we coded for both of these variables. The following characteristics were also chosen, as we believed they are important sample-specific factors that might influence score reliability estimates and many of them are commonly examined study characteristics included in RG studies (e.g., Lopez-Pina et al., 2009; Kieffer & MacDonald, 2011). The variables included: (a) Age range of participants (coded as a continuous variable), (b) Mean age of participants (coded as a continuous variable), (c) Type of sample (coded as 1 for clinical sample, 2 for community sample and 3 for residential sample), (d) Presence of medical or mental disorders (coded as 1 for specific medical condition, 2 for specific mental disorders and 3 for both), (e) Sample size (coded as a continuous variable), (f) Age standard deviation (coded as a continuous variable), (g) Mean test score (coded as a continuous variable), and (h) language (coded as English versus other language). We also noted the publication year to examine the degree of change in score reliabilities over time. In order to examine the reliability of the coding process the second author coded 43 of the articles (20% of the sample) coded by the first author and the mean agreement was .95. Discrepancies in coding were resolved through discussions between the raters and a review of the study. It is important to note that sample characteristics were reported for each reliability value derived from a study. For example, if the coefficients were reported separately for men and women we also coded the study characteristics separately for men and women.

Results

Prior to using the reliability coefficient in the RG analysis, we transformed them into a more amenable form. Reliability coefficients such as Cronbach's alpha are often considered as r^2-equivalent as they are variance accounted-for statistics (Thompson & Vacha-Hasse, 2000). Therefore, we first took the square roots of the reliability coefficients to convert them into the same metric as a correlation. Then, following standard meta-analytic methodology (e.g. Feldt & Charter, 2006; Thompson & Vacha-Haase, 2000), we applied Fisher's *r*-to-*z* transformation to correct the fact that correlations are not normally distributed and have a tendency to be skewed. Following our analyses, the transformation process was reversed so that we could report means and confidence intervals in the same form as the original Cronbach's alpha and kappa values. All analyses used reliability coefficients weighted by the inverse variance weight (*n*-3; Graham et al., 2006) and were conducted with Wilson's (2005) random effects meta-analysis macros for SPSS. This ensured that more weight was given to reliability estimates obtained from studies that had large sample sizes, as they are more likely to yield results that are similar to the true population values. Table 2 shows the descriptive statistics for each of the twelve anxiety measures. Wilson's (2005) macros were used to calculate the mean reliability coefficient produced by each measure as well as the 95% confidence intervals. Table 2 also presents the highest and lowest reliability estimate reported in the reviewed articles as well as the number of reliability estimate reported in the literature for each measure (*k*).

Q statistics and I^2 index

As conducted in previous RG (e.g. Graham & al., 2011; Lopez-Pina et al., 2009), we used *Q* test of homogeneity to assess the degree of dispersion of reliability estimates around the mean reliability coefficients for each measure. In other words, we examined whether the reliability estimates are situated around the mean or if they are more dispersed than would be expected if all

the values were derived from the same population. The I^2 index was also calculated to compliment the results of the *Q* test. The I^2 index represents the percentage of the total variability in a set of reliability estimates that is caused by true heterogeneity. As reported in Table 2, most measures (BAI, GADS, HADS, PSWQ, STAI, SCL-90-R and WS) showed a statistically significant *Q* test suggesting that the variability seen in the reliability estimates cannot be due to sampling error alone. Because measures with significant *Q* statistics and high I^2 coefficients have noteworthy variance in reliability coefficients, it is appropriate to examine which sample and study characteristics might best explain this variability. It should be noted that as the GHQ had only one reported alpha coefficient and no studies reported an alpha coefficient for the HARS, it was impossible to calculate the *Q* statistics for these measures. Additionally, three measures (BSI, GAI, GMSE) showed a non-significant *Q* statistic suggesting that these reliability estimates are homogeneous and that the measures do not tend to produce different reliability estimates with each administration. It is however likely that the low number of usable reliability coefficient (*k*) might have been responsible for the lack of significance of the *Q* statistics. Therefore, we cannot make definitive statements about the homogeneity of reliability estimates produced by these three measures until further data is available.

Orwin's fail-safe N

As a large number of study failed to report reliability coefficients, it is important to examine the possible influence missing reliability coefficients might have on the mean reliability values we obtained. It is possible that the reliability of scores in studies who did not report a reliability estimate were different from those included in the RG. To this end, we used Orwin's (1983) fail-safe *N* calculation. This method indicates how many studies with an average reliability of .5 would be needed to bring down the mean reliability below a cutoff of .7. As seen in Table 2, few studies with a reliability coefficient below .5 would be needed for some measures

to bring the average reliability below the cutoff of .7. For example, 7 studies presenting reliability coefficients below.5 would be necessary to bring the mean reliability of the BAI (α = .86) below the cutoff of .7. Such values suggest that the results of the RG should be seen as a tentative, but best estimate of reliability until more studies that report usable reliability estimates are collected, as low fail-safe *N* values are often found when the number of available reliability values is limited (e.g., Graham & Christiansen, 2009). These values are significantly lower than what has been found in other studies with larger sample sizes (e.g., Graham et al., 2000), where Orwin's statistic suggests that hundreds of unpublished articles with an average reliability coefficient below .5 would be needed to bring the mean reliability of measures below .7.

Analysis of Moderator Variables

We conducted a series of bivariate correlations using Wilson's (2005) macro's in SPSS to examine the sample and study characteristics that predicted score reliabilities. Although previous RG studies have often used multiple regressions, our sample size was small and therefore more appropriate for correlational analyses, which have also been used in RG studies (e.g. Graham et al., 2011). We first calculated a series of bivariate correlations to test the relations between all the possible moderators and the mean reliability coefficients of all measures combined. As seen in Table 3, six of all the possible moderators (mean age, age *SD*, sample size, proportion of men, score *SD* and number of items) showed a statistically positive significant relationship with the reliability estimates. As noted in previous RG studies and consistent with the principles of classical test theory, the reliability of scores obtained on the anxiety measures increased as the number of items of the measures, the variability of test scores and the sample size of the population increased. Researchers and clinicians should also keep in mind that the reliability of the anxiety measures tend to be higher in samples with a higher age mean and a higher proportion of men.

We then conducted a series of bivariate correlations to test the relations between the six significant moderators and the mean reliability of test scores of the individual measures. Ideally, such analyses would be conducted for each individual measure, however because of the limited number of studies reporting reliability values only two measures (HADS and STAI) had sufficient reliability coefficients (*k*) to conduct such analyses. The results are reported in Table 4 and are described further in the section. As each correlation presented in Table 3 and Table 4 has a different number of reliability coefficients, it is important to not only look at statistical significance but also at the magnitude of the correlation as preliminary evidence of the relation between the variable and the reliability coefficients.

Measures Demonstrating Excellent Mean Reliability Scores

Internal consistency can be considered as adequate when a preponderance of evidence indicates an alpha value of .70-.79, good when the alpha is between .80-.89, and excellent when the alpha is above .90 (Hunsley & Mash, 2008). According to this guideline, three measures showed an excellent mean reliability score with alpha values of .9 and above. The reliability of the GAI was the highest of the measures of anxiety in the present meta-analysis at .92 and a 95% confidence interval ranging from .9 to .93. As the average drew from only 3 reliability coefficients, it is not surprising that there was not a sufficient amount of variance in GAI reliabilities to result in a significant *Q* and therefore, we did not conduct any moderator analyses on this measure. The GAI has not been widely used in the literature as it was recently developed and has to compete against many other established measures. However, although limited information is available regarding the utility of the GAI, the results of the current RG and the fact that it was specifically created for an older population suggests it has the potential to be a reliable measure to assess the presence of anxiety symptoms in older adults.

Both the PSWQ and the SCL-90-R had an average reliability score of .9. The 95% confidence interval ranged from .84 to.94 for the PSWQ and from .78 to .95 for the SCL-90-R. Although, only 4 studies contributed to both of these averages, the variance between those reliability coefficients was sufficiently high to result in a statistically significant *Q*. However, because of the small *k* of both measures, we could not conduct further analysis to determinate whether possible moderator variables could explain part of the variability. There is some concern that some older adults have difficulty completing and interpreting the content of the reversed items of the PSWQ (Stanley et al., 2003; Wetherell et al., 2003), but the results of the present RG suggest it has good internal consistency and that it has the potential to be a useful measure when assessing the presence of worry in leer adults. When it comes to the SCL-90-R, there is some evidence that most subscales do not adequately measure the constructs they are designed to measure and that it tends to overpathologize (Hunsley & Lee, 2010). Although it demonstrated high reliability in the present RG, the SCL-90-R might be best used as a brief measure of general psychological distress rather than only using the anxiety-related subscales.

Measures Demonstrating Good Mean Reliability Scores

Four measures showed a good mean reliability score with alpha values ranging from .8 to .89. Both the BAI and the WS had an average reliability score of .86. The 95% confidence interval ranged from .82 to .86 for the BAI and from .83 to .86 for the WS. Although few studies contributed to both averages, the variance between the reliability coefficients was sufficiently high to result in a statistically significant *Q*. However, because of the small *k* of both measures, we could not conduct further analyses to determinate whether possible moderator variables could explain part of the variability. Overall, the reliability of the BAI as well as its simplicity suggests it might be a useful tool to detect the presence of anxiety in older adults. However, because of potential confounds with depressive symptoms and high somatic item content (13 of the 21 items

are related to somatic symptoms), we encourage researchers and clinicians to be vigilant when using the BAI in a medical setting or with older adults presenting with a medical condition. On the other hand, the results of the present RG and the fact that the WS was specifically created for an older population suggest it has the potential to be a useful measure of worry in older adults.

The STAI had an average reliability of .82 and a 95% interval confidence ranging from .7 to .89. The variance between the reliability coefficients was sufficiently high to result in a statistically significant Q and the I^2 statistics suggest that 97 % of the total variability among reliability estimates is due to true heterogeneity among studies. We conducted a correlational analysis predicting score reliabilities with the sample size, the proportion of men and the mean age. Other variables could not be included in the analysis as they were infrequently reported in the studies that used the STAI in the RG. Specifically researchers and clinicians need to keep in mind that the STAI tended to produce more reliable scores with larger sample sizes and in samples consisting of a higher proportion of men. Although the STAI is lengthy and can be easily misinterpreted by older adults (Dennis et al., 2007), the results of the present RG suggest it has good internal consistency and that it has the potential to be a useful measure when assessing anxiety in older adults.

The BSI had an average reliability of .81 and a 95% confidence interval ranging from .76 and .85. Considering that the average reliability was computed from only 4 reliability coefficients, it is not surprising that the variance between the coefficients was not sufficiently high to result in a statistically significant Q. Because of the small k and the Q statistic indicating that insufficient variability exists among the alpha estimates, we did not proceed with further analysis of the moderator variables. Like the SCL-90-R, on which it is based, researchers and clinicians might consider the BSI to assess general distress rather than anxiety symptoms specifically.

Measures Demonstrating Adequate Mean Reliability Scores

Three measures showed an adequate mean reliability score with alpha values ranging from .7 to .79. The HADS had an average reliability of .79 and a 95% confidence interval ranging from .77 to .81. The variance between the reliability coefficients was sufficiently high to result in a statistically significant Q and the I^2 statistics suggest that 59% of the total variability among reliability estimates is due to true heterogeneity among studies. We conducted a correlational analysis predicting score reliabilities with the mean age of the sample, the standard age deviation, the proportion of men and the sample size. Researchers and clinicians should keep in mind that the HADS produces more reliable scores in larger samples as well as in samples with a higher mean age. As the measure is unbiased by comorbid medical conditions, researchers and clinicians might consider the HADS when working in a medical setting or with older adults presenting with a medical condition.

The GADS had an average reliability of .74 and a 95% confidence interval ranging from .66 to .81. Given that the GADS had the lowest mean alpha coefficient in the present meta-analysis and that the lower confidence bound interval is lower than .70, the reliability of the GADS appears to be problematic relative to other available anxiety measures. The variance between the 6 reliability coefficients was sufficiently high to result in a statistically significant Q. However, due to the limited sample size we did not pursue further analysis to calculate the possible effects of moderating variables on the variability of reliability between studies. The GADS include somatic symptoms (e.g., waking early, headaches) that may overestimate the prevalence of anxiety and depression in older adults resulting in a high rate of classification errors. It is likely not a good first choice when choosing an anxiety measure for an older population.

As the GMSE is a semi-structured clinical interview, we examined the inter-rater reliability of the measure through kappa. The mean reliability of the GMSE should therefore not be compared with other measures. The mean inter-rater reliability of the GMSE was .72 with a 95% confidence interval ranging from .66 to .77. Given that the lower confidence bound interval is below .7, the reliability of the GMSE appears to be problematic. The variance between the 5 reliability coefficients was not sufficiently high to result in a statistically significant *Q*. Because of the small *k* and the fact that the *Q* statistic indicated that insufficient variability exists among the alpha estimates, we did not proceed with further analysis of the moderator variables. The strength of the GMSE resides in the fact that it excludes the effects of physical illness, which can be a useful when assessing the mental health of older adults in medical settings. However, as the GMSE has not been used in research published in the last decade and has shown relatively low inter-rater reliability in the present RG, it is likely not a good first choice when choosing an anxiety measure for an older population.

Measures with Insufficient Reliability Information

There were two measures for which there was insufficient information for them to be included in the present RG. First, no studies in the present RG reported an alpha coefficient for the HARS and its usefulness with older adults has been questioned due to the heavy emphasis placed on somatic symptoms (e.g., tension) that are common experiences in aging individuals (Kogan et al., 2000; Skopp et al., 2006). Additionally, although the GHQ is commonly used to detect the presence of mental disorders, very little is known about its use with older adults and there was only one study that reported an alpha coefficient ($\alpha = .82$) for the GHQ. We cannot make any definitive statements about the utility of the HARS or the GHQ with such limited information and encourage researchers and clinicians to be cautious when using one of these measures with an older population.

Discussion

The purpose of the present study was to explore the typical score reliability of the anxiety measures most commonly used with older adults and to identify study-specific factors that might have an impact on the reliability estimates of the measures. A RG study was conducted on ten of the twelve anxiety measures most commonly used in research with older adults: BAI, BSI, GAI, GADS, GMSE, HADS, PSWQ, the trait version of the STAI, SCL-90-R, and the WS as only one study reported a reliability coefficient for the GHQ, and none for the HARS, we could not conduct analyses on these measures. Of the measures examined in this study, the GAI, BAI, PSWQ and WS showed sufficiently strong reliability evidence to warrant their use when assessing anxiety in older adults. These results are in line with the findings of a systematic review conducted by the authors (Therrien & Hunsley, 2012) in which the four measures also showed sufficiently strong psychometric properties to justify their use with older adults. This suggests that based on their overall level of reliability and previous psychometric evidence, both researchers and clinicians assessing anxiety in a geriatric population should consider these measures as likely to be the best currently available. The SCL-90-R, the trait version of the STAI, BSI and HADS all appear to be adequate measures of anxiety in older adults, at least from the perspective of score reliability. However, these measures have limitations in other psychometric parameters that should be considered by researchers (Therrien & Hunsley, 2012). The present data suggest that the GADS and the GMSE are weak alternatives for measuring of anxiety in older adults. Although average scores of reliability were marginally acceptable for these measures, they were the lowest of all those examined here. Finally, much more information on the psychometric properties of the GHQ and the HARS is necessary before any sort of decision can be made regarding their value in assessing anxiety in older adults.

Considerable effort has been made to develop anxiety measures specifically suited for older adults and, based on this RG, two (GAI and WS) of these three measures showed high reliability estimates that warrant their use in assessing anxiety in older adults. In many instances, these instruments should be preferred over more commonly used measures that were not developed for older populations and whose psychometric properties are either low or largely unknown with regards to older adults. Measures created specifically for older adults such as the GAI and the WS have many advantages over measures developed for younger adults including superior psychometric properties, norms established for an older population and increased acceptability among older adults (Edelstein & Segal, 2010), Additionally, they can address specific late life issues such as how developmental and medical factors influence anxiety symptoms.

The Effect of Moderating Variables on the Reliability of Anxiety Measures

Given the limited number of studies that reported reliability estimates for their data, we had very limited opportunities to examine moderator variables despite evidence of significant variability in the reliability estimates of most measures. However, we were able to combine the information from all measures and conduct bivariate correlations between score reliability estimates and the possible moderator variables. As expected from psychometric theory, the variability of the test scores of a sample and the number of items in the measure were positively associated with reliability estimates. Indeed, this is not surprising as research indicates that measures with greater variance and greater numbers of test items normally generate larger and more stable reliability coefficients (Kieffer & MacDonald., 2011). Additionally, measures showed higher reliability values in larger samples, as well as with older individuals (mean age and age standard deviation), and with samples presenting with a higher proportion of men. It is important to consider that although statistically significant, the magnitude of each of these correlations was relatively small and was based on a limited number of studies.

Seven of the measures (BAI, GADS, HADS, PSWQ, the trait version of the STAI, SCL-90-R and WS) had noteworthy and significant variance in their reliability coefficients. Although it would have been interesting to examine which study characteristics accounted for this variability, only the HADS and the trait version of the STAI had sufficient reported reliability values to pursue such analyses. In general the results suggested that the HADS produced higher reliability scores in samples consisting with a higher mean age, and the STAI produced higher reliability scores in larger samples and in samples consisting of a higher proportion of men. The considerable heterogeneity we found for the other anxiety instruments can only be explored once more reliability values (i.e., more studies) are available. In the meantime these results suggest that researchers and clinicians must be aware that some study or sample specific characteristics such as the age and gender distribution of the sample might influence the reliability of the anxiety measures they choose to use.

Limited Reporting of Reliability Estimates

The tendency to not report reliability coefficients for study data seems to have prevailed despite the guidelines published by various organizations (e.g., Wilkinson and the American Psychological Association Task Force on Statistical Inference, 1999). In the present RG, an astounding 76% of the reviewed studies either made no mention of reliability or merely reported inducted values as if they applied to their data by referencing the test manual or a previous study. As emphasized by Vacha-Haase et al. (2000), researchers using reliability estimates from previous research as a way to demonstrate reliability in the current data must ensure that sample characteristics and variability within the samples are comparable. Because every sample has unique characteristics that could lead to different set of scores on a measure, and because every set of scores has a unique internal consistency reliability coefficient, it cannot be assumed that the reliability presented in test manuals will be equal or even similar to that obtained in the sample at

hand. In addition, several researchers have found that the score reliability estimates varies significantly on different administration of the same instrument (e.g., Caruso, 2000). In our RG study, most of the commonly used measures for older adults were created for younger adults and inducting reliability form test manuals is simply not appropriate. The best practice, undoubtedly, is to report score reliability estimates for the data used in a study.

Limitations of the Study

A few limitations of the present study should be noted. First, as with all meta-analyses, the main limitation resides in the ability to identify and retrieve all of the studies that have used an anxiety measure with a sample of older adults. We searched for studies in the most important databases for psychology (PsycInfo) and medicine (PubMed) but other databases were not considered. Our search strategy did ensure that we considered the majority of journals that publish research on the health or mental health of older adults. However, as it is likely that only a subset of all studies that use an anxiety measure are published in the literature, it is possible that unpublished studies might have lower reliability coefficient than published ones, which would consequently modify our obtained mean reliability values.

Second, as is the case with RG studies, very few studies included reliability values for the anxiety measures of interest. As such, our analyses and results are based on the studies that reported adequate information about their samples and do not consider all the studies that used an anxiety measure with older adults. Therefore, when interpreted in this context, our results should be considered as the currently available best estimate of the mean reliability values for each measure.

Conclusion

Many measures of anxiety have been developed but, when it comes to assessing anxiety in older adults, no clear consensus exist among researchers on which measures are psychometrically

strongest. Of the measures examined in this study, the GAI, BAI, PSWQ and WS showed sufficiently strong reliability evidence to warrant their use when assessing anxiety in older adults. Although reliability is important when selecting and using a measure, a psychometrically strong measure should also have evidence of validity (discriminant and concurrent) and clinical utility. It should also possess appropriate norms for norm-referenced interpretation and replicated supporting evidence for accuracy of cut-scores. Although the recommendations made in the present study are primarily based on reliability, we encourage researchers and clinicians to consider the overall evidence of psychometric quality when choosing an anxiety measure for older adults. Considering the limited research on the psychometric properties of anxiety measures for older adults, clinicians and researchers are advised to be cautious and to carefully consider the strengths and weaknesses of each measure before deciding which one to use for a specific purpose. We also encourage researchers to routinely calculate and report reliability for their own set of data instead of relying on reliability induction, even if the study is not a psychometric or measure validation study. As found in most RG studies, there is evidence of limited reporting of reliability scores and, therefore, a need for improvement in reporting practices. Additionally, given the utility and increasing use of meta-analysis, we encourage researchers to be thorough when describing the samples in their study and to routinely report characteristics such as the mean age and age variance for a sample, along with the mean and standard deviation for each measure used in the study.

References

Alwahhabi, F. (2003). Anxiety symptoms and generalized anxiety disorder in the elderly: A review. *Harvard Review of Psychiatry, 11*, 180-193.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Barnes, L. L. B., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement, 62*, 603-618.

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology, 56,* 893-897.

Brenes, G. A., Kritchevsky S. B., Mehta, K. M., Yaffe, K., Simonsick, E. M., Ayonayon, H.N et al. (2007). Scared to death: Results from the Health, Aging and Body Composition study. *American Journal of Geriatric Psychiatry, 15*, 262-265.

Bryant, C., Jackson, H., & Ames, D. (2009). Depression and anxiety in medically unwell older adults: Prevalence and short-term course. *International Psychogeriatrics, 21*, 754-763.

Charter, R. A. (2007). A practical use for the KR-21 reliability coefficient. *Psychological Reports, 101*, 673-674.

Copeland, J. R. M., Kelleher, M. J., Kellett, J. M., Gourlay, A. J., Gurland, B. J., Fleiss, & Sharpe, L. (1976). A semi-structured clinical interview for the assessment of diagnosis and mental state in the elderly: The Geriatric Mental State Schedule I: Development and reliability. *Psychological Medicine, 6*, 439-449.

Dennis, R. E., Boddington, S. J., & Funnell, N. J. (2007). Self-report measures of anxiety: Are they suitable for older adults? *Aging & Mental Health, 11,* 668-677.

Derogatis, L. R. (1994). *Administration, scoring and procedures manual*. Minneapolis, MN: National Computer System.

Derogatis, L. R., & Spencer, P. M. (1982). *The Brief Symptom Inventory (BSI): Administration, and procedures manual-I*. Baltimore, MD: Clinical Psychometric Research.

Feldt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement, 66*, 215-227.

Gladstone, G. L., Parker, G. B., Mitchell, P. B., Malhi, G. S., Wilhelm, K. A., & Austin, M. P. (2005). A brief measure of worry severity (BMWS): Personality and clinical correlates of severe worriers. *Journal of Anxiety Disorders, 19*, 877-892.

Goldberg D. et al. (1978). *Manual of the General Health Questionnaire*. Windsor, England: NFER Publishing.

Graham, J. M. & Christiansen, K. (2009). The reliability of romantic love: A reliability generalization meta-analysis. *Personal Relationship, 16*, 49-66.

Graham, J. M., Diebels, K. J, & Barnow, Z.B. (2011). The reliability of relationship satisfaction: A reliability generalization meta-analysis. *Journal of Family Psychology, 25*, 39-48.

Graham, J. M., Yenling, J. L. & Jeziorski, J. L. (2006). The Dyadic Adjustment Scale: A reliability generalization meta-analysis. *Journal of Marriage and Family, 68*, 701-717.

Hamilton, M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology, 32*, 50-55.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality Schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology, 10*, 249-254.

Henson, R., Kogan, L. & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement*, *61*, 404-420.

Holi, M. (2003). *Assessment of psychiatric symptoms using the SCL-90-R.* (Unpublished doctoral dissertation). Helsinki University, Finland.

Hunsley, J. & Lee, C. (2010). *Introduction to clinical psychology: An evidence based approach* (2nd Ed.). Mississauga, ON: John Wiley & Sons Canada.

Hunsley, J. & Mash E. J. (2008). *A guide to assessment that works.* New York: Oxford University Press.

Kieffer, K. M. & Reese, R. J. (2002). A reliability generalization study of the Geriatric Depression Scale (GDS). *Educational and Psychological Measurement, 62*, 969-994.

Kieffer, K. M., & MacDonald, G. (2011). Exploring factors that affect score reliability and variability in the Ways of Coping Questionnaire reliability coefficients: A meta-analytic reliability generalization study. *Journal of Individual Differences, 32(1)*, 26-38.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Kvall, K., Macijauskiene, J., Engedal, K., & Laake, K. (2001). High prevalence of anxiety symptoms in hospitalized geriatric patients. *International Journal of Geriatric Psychiatry, 16*, 690-693.

Kvaal, K., McDougall, F., Brayne, C., Matthews, F., & Dewey, M. (2008). Co-occurrence of anxiety and depressive disorders in a community sample of older people: results from the MRCCFAS. *International Journal of Geriatric Psychiatry, 23*, 229-237.

Lane, G. G, White, A. E, & Henson, R. K. (2002). Expanding reliability generalization methods with KR-21 estimates: An RG study of the Coopersmith Self-Esteem Inventory. *Educational and Psychological Measurement*, *62*, 685-711.

Lopez-Pina, J., Sanchez-Meca, J., & Rosa-Alcazar, A. (2009). The Hamilton Rating Scale for Depression : A meta-analytic reliability generalization study. *International Journal of Clinical and Health Psychology*, *9*, 143-159.

Orwin, R.G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*, 157–159.

Pachana, N. A., Byrne, G. J., Siddle, H., Koloski, N., Harley, E., & Arnold, E. (2007). Development and validation of the Geriatric Anxiety Inventory. *International Psychogeriatrics, 19*, 103-114.

Preisig, M., Merikangas K. R., & Angst, J. (2001). Clinical significance and comorbidity of subthreshold depression and anxiety in the community. *Acta Psychiatrica Scandinavia,104*, 96-103.

Rousse, S. V. (2007). Using reliability generalization methods to explore Measurement error: An Illustration using the MMPI-2 PSY-5 scales. *Journal of Personality Assessment*, *88*, 264-275.

Segal, D. L., June, A., Payne, M., Coolidge, F. L., & Yochim, B. (2010). Development and initial validation of a self-report assessment tool for anxiety among older adults: The Geriatric Anxiety Scale. *Journal of Anxiety Disorders*, *24*, 709-714.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory STAI (Form Y): Self-evaluation questionnaire*. Palo Alto, CA: Consulting Psychologists Press.

Stanley, M. A., Beck, J. G., Novy, D. M., Averill, P. M., Swann, A. C., Diefenbach, G. J., & Hopko, D. R. (2003). Cognitive-behavioral treatment of late-life generalized anxiety disorder. *Journal of Consulting and Clinical Psychology, 71*, 309-319.

Therrien, Z., & Hunsley, J. (2012). Assessment of anxiety in older adults: A systematic review of commonly used measures. *Aging and Mental Health, 16,* 1-16.

Thompson, B. & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60(2)*, 174-195.

Twiss, E., Seaver, J., & McCaffrey, R. (2006). The effect of music listening on older adults, *Nursing in Critical Care, 11*, 224-231.

Vacha-Haase, T. (1998). Reliability generalization exploring variance in measurement error affecting score reliability across studies. *Educational & Psychological Measurement*, *58*, 6-20.

Vacha-Haase, T., Henson, R., & Caruso, J. (2002). Reliability Generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*, 562-569.

Vacha-Haase, T., Kogan, L. R. & Thompson, B. (2000). Sample composition and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*, 509–552.

Vacha-Haase, T., Ness, C., Nilsson, J. & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *Journal of Experimental Education, 67(4)*, 335-341.

Van Zelst, W. H., De Beurs, E., Beekman, A. T. F., Van Dyck, R., & Deeg, D. (2006). Well-being, physical functioning and use of health services in the elderly with PTSD and subthreshold PTSD. *International Geriatric Psychiatry, 21,*180-188.

Vassar, M., & Bradley, G. (2010). A reliability generalization study of coefficient alpha for the Life Orientation Test. *Journal of Personality Assessment, 92*, 362-370.

Vassar, M. & Crosby, W. (2008). A reliability generalization study of coefficient alpha for the UCLA Loneliness Scale. *Journal of Personality Assessment*, *90*, 601-607.

Wheeler, D. L., Vassar, M., Worley, J. A., & Barnes, L. L. B. (2011). A reliability generalization meta-analysis of coefficient alpha for the Maslach Burnout Inventory. *Educational and Psychological Measurement, 71*, 231-244.

Wetherell, J. L., Gatz, M., & Craske, M. G. (2003). Treatment of generalized anxiety disorder in older adults. *Journal of Consulting and Clinical Psychology, 71*, 31-40.

Wilson, D. B. (2005). Meta-analysis macros for SAS, SPSS and Stata. Retrieved November 2011, from heep://mason.gmu.edu~dwilsonb/ma.html.

Wisocki, P. A., Handen, B., & Morse, C. K. (1986). The Worry Scale as a measure of anxiety among homebound and community active elderly. *The Behavior Therapist, 9,* 91-95.

Wolitzky-Taylor, K. B., Castriotta, N., Lenze, E. J., Stanley, M. A. & Craske, M. G. (2010). Anxiety disorders in older adults: A comprehensive review. *Depression and Anxiety, 27*, 190-211.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression scale. *Acta Psychiatrica Scandinavica, 67*, 361-370.

Table 1

Most commonly used anxiety measures as found by Therrien & Hunsley (2012)

Measure	Number of times used in literature	Number of items	Type of anxiety being assessed	Scoring	Intended population
Beck Anxiety Inventory (BAI)	8	21-item self-report questionnaire.	General measure of anxiety.	Four-point scale with scores ranging from 0 to 63.	Young and middle aged adults.
Brief Symptom Inventory (BSI)	7	53-item self-report questionnaire.	General measure of anxiety and phobic anxiety. Also includes other subscales: somatization, obsessive-compulsive, interpersonal sensitivity, depression, hostility, paranoid ideation and psychotism.	Five-point scale with scores ranging from 0 to 72.	Adolescents, young and middle aged adults.
Geriatric Anxiety Inventory (GAI)	4	20-item self-report questionnaire.	General measure of anxiety.	Agree or disagree answers with scores ranging from 0 to 20.	Older adults.
Goldberg Anxiety and Depression Scale (GADS)	8	18-item self-report questionnaire.	General measure of anxiety. Also includes a depression subscale.	Yes (1) or no (0) answers with scores ranging from 0 to 9 for each	Young and middle aged adults

				scale.	
General Health Questionnaire (GHQ)	8	28-item self report questionnaire.	General measure of anxiety. Also includes other subscales: somatic symptoms, social dysfunction and severe depression.	Four-point scale with scores ranging from 0 to 84	Adolescents, young and middle aged adults.
Geriatric Mental State Examination (GMSE)	14	541-item semi-structured interview.	General measure of anxiety. Also includes other subscales: organicity, schizophrenia, mania, depression, hypocondrias and phobias.	Diagnostic confidence level for each subscale ranging from 0 (no symptoms) to 5 (very severely impaired).	Older adults.
Hospital Anxiety and Depression Scale (HADS)	26	14-item self report questionnaire.	General measure of anxiety. Also includes a depression subscale.	Four-point scale with scores ranging from 0 to 21.	Medical outpatients aged between 16 and 65.
Hamilton Anxiety Rating Scale (HARS)	13	14-item clinician-administered rating scale.	Measure of psychic/cognitive anxiety and somatic anxiety.	Five-point scale with scores ranging from 0 to 56.	Young and middle aged adults.
Penn State Worry Questionnaire (PSWQ)	6	16-item self-report questionnaire.	Measure of pathological worry.	Five-point scale with scores	Young and middle aged adults.

				ranging from 16 to 80.	
Symptom Chelist-90-Revised (SCL-90-R)	6	90-item self-report questionnaire.	General measure of anxiety and phobic anxiety. Also includes other subscales: somatization, obsessive-compulsive, interpersonal sensitivity, depression, hostility, paranoid ideation and psychotism.	Five-point scale with scores ranging from 0 to 72.	Adolescents, young and middle aged adults.
State Trait Anxiety Inventory (STAI)	35	40-item self report questionnaire	Measure of state (temporary condition) and trait (long standing) anxiety.	Four-point scale with scores ranging from 20 to 80.	Young and middle aged adults
Worry Scale (WS)	5	35-item self-report questionnaire.	Measure of pathological worry.	Five-point scale with score ranging from 0 to 140.	Older adults.

Table 2
Descriptive Statistics for Anxiety Scales' Score Reliabilities

Measure	*k*	Mean α	*95% confidence interval* Lower	Upper	Min.	Max.	*Q*	I^2	Orwin's *N*
BAI	3	.86	.82	.86	.78	.92	17.74*	88.72	7
BSI	4	.81	.76	.85	.74	.85	3.43	12.47	8
GAI	3	.92	.90	.93	.90	.93	1.91	0.00	10
GADS	6	.74	.66	.81	.26	.85	134.69*	96.29	13
GHQ	1	.82	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GMSE	5	.72**	.66	.77	.64	.78	3.76	0.00	6
HADS	20	.79	.77	.81	.72	.85	58.91*	69.44	24
HARS	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
PSWQ	4	.90	.84	.94	.79	.95	54.54*	87.78	13
STAI	12	.82	.70	.89	.32	.92	229.10*	95.20	22
SCL-90-R	4	.90	.78	.95	.82	.96	140.12*	97.86	13
WS	2	.86	.83	.89	.82	.90	5.98*	66.54	8

Note. Min. = Minimum reliability coefficient reported in the literature; Max. = Maximum reliability coefficient reported in the literature; *k* = number of reliability coefficients presented in the reviewed studies; BAI = Beck Anxiety Inventory; BSI = Brief Symptom Inventory; GAI = Geriatric Anxiety Inventory; GADS = Goldberg Anxiety and Depression Scale; GHQ = General Health Questionnaire; GMSE = Geriatric Mental Status Examination; HADS = Hospital Anxiety and Depression Scale; HARS = Hamilton Anxiety Rating Scale; PSWQ= Penn State Worry Questionnaire; STAI = State Trait Anxiety Inventory; SCL-90-R = Symptom Checklist-90-Revised; WS = Worry Scale.
* $p \leq 0.05$, ** Inter-rater (Kappa values)

Table 3

Bivariate Random-Effects Weighted Maximum-Likelihood Correlation (and k values) Between Score Reliability and Sample Characteristics Matrix For All Measures Combined

Mean age	**Age SD**	**Medical sample**	**Mental health sample**	**Community sample**	**Sample size**
.13* (45)	.13* (37)	.05 (58)	.03 (58)	-.06 (57)	.13* (58)
Proportion of men	**Year**	**Mean Score**	**Score SD**	**Translation**	**Number of items**
.09* (55)	-.09 (58)	-.09 (16)	.15* (14)	.03 (57)	.46* (57)

Note. Numbers in parentheses are the number of data points in each correlation. * $p \leq 0.05$

Table 4

Bivariate Random-Effects Weighted Maximum-Likelihood Correlation (and k values) Between Score Reliability and Significant Sample Characteristics Matrix For the HADS and the STAI

Measure	Mean Age	Age SD	Sample size	Proportion of men
HADS	.18 (13)	.10 (12)	.46* (19)	-.14 (18)
STAI	.44 (7)	--	.21* (12)	.25* (11)

Note. HADS = Hospital Anxiety and Depression Scale; STAI= State Trait Anxiety Inventory.
Numbers in parentheses are the number of data points in each correlation. Blank cells indicate an insufficient number of studies to calculate a correlation. * $p \leq 0.05$

Comparing Approaches for Categorizing Measure Reliability in the Assessment of Anxiety in Older Adults *

Zoé Therrien John Hunsley

University of Ottawa, Canada

* This study has been submitted for publication at *Administration and Policy in Mental Health Services Research*.

Abstract

A systematic review of the anxiety measures most commonly used with older adults found that most are developed for a younger population and lacked sufficient evidence of their psychometric properties when used with an older population. As researchers and clinicians need a fast way to judge whether a measure is appropriate when faced with daily assessment tasks, recent efforts have been made in developing guidelines to operationalize the criteria necessary to designate a measure as evidence-based. The goal of this study is to: a) apply the reliability criteria for two existing evidence based assessment categorization systems to the most commonly used anxiety measures with older adults and b) compare the results of these categorization systems to the results obtained in a previous reliability generalization study that gave the mean reliability of each measure. We discuss the strengths and limitations of both approaches and suggest how researchers and clinicians can identify psychometrically sound measures without having to conduct more labor-intensive meta-analysis studies.

Keywords: older adults, anxiety, evidence-based assessment, evidence-based instruments psychometric properties,

Comparing Approaches for Categorizing Measure Reliability in the Assessment of Anxiety in Older Adults

The rich history of assessment within clinical psychology has left the field with a confidence regarding the value of existing psychological assessment methods. An unfortunate consequence of this is that assessment procedures are rarely questioned or evaluated. This is especially problematic, as research suggests that many commonly used assessment methods and instruments are not supported by scientific evidence (e.g., Hunsley, Lee, & Wood, 2003; Hunsley & Mash, 2007; Norcross, Koocher, & Garofalo, 2006). This tendency to use instruments without closely examining their scientific qualities may also reflect the influence of other interrelated factors such as a belief in the intrinsic worth of many commonly used assessment measures, a lack of information regarding the evidence needed to evaluate the utility of assessment measures and a greater interest in treatment and intervention activities among psychologists and researchers (Mash & Hunsley, 2005).

The tendency to simply assume that a measure is psychometrically sound is evident with respect to the assessment instruments used with older adults. A recent systematic review showed that the anxiety measures most commonly used with older adults are mostly measures developed for a younger population and that the majority lack sufficient evidence of their psychometric properties when used with an older population (Therrien & Hunsley, 2012a). In order to provide guidance to both clinicians and researchers, the primary goal of this study is to examine whether the anxiety measures used with older adults can be consistently and accurately categorized as evidence-based and thus appropriate for use. As an initial step, this will be done by applying the reliability criteria of existing evidence-based assessment categorization systems and comparing the results of these applications to the results of a reliability meta-analysis conducted with the anxiety measures.

Evidence-Based Assessment (EBA) and Instrument Categorization Approaches

As part of the evidence-based practice (EBP) movement, a modest but noticeable shift has occurred in the assessment literature and more attention is being paid to the way in which psychological assessment is conducted and used to guide practice. For example, clinicians are seeking assessment tools they can use to determinate a client's level of pretreatment functioning and to develop, monitor and evaluate the services received by the client (Barkham et al., 2001; Hatfield & Ogles, 2004). Such changes have led to the development of evidence-based assessment (EBA), an approach to clinical evaluation that emphasizes the use of research and theory to guide the selection of constructs, the methods and measures used in the assessment as well as the overall assessment process (Hunsley & Mash, 2007).

Evaluating whether an assessment measure is psychometrically sound can be an arduous task, as psychometric properties are not properties of an instrument *per se* but rather are properties of an instrument when used with a specific sample for a specific purpose (Barnes, Harp & Jung, 2002; Henson, Kogan & Vacha-Haase, 2001). In other words, as reliability is affected by the composition and variability of the sample, a measure can produce reliable scores in one study and unreliable scores in another. For this reason, psychometricians and test developers have been reluctant to elaborate precise standards for psychometric properties that an instrument must meet in order to be considered useful for different assessment purposes (e.g., Streiner & Norman, 2008). Regardless, researchers and clinicians constantly need to decide whether an instrument is good enough for the assessment task at hand and, without specific guidelines, there is a risk that they might turn to measures that have little or no supporting psychometric evidence. Fortunately, recent efforts have been made in developing guidelines to operationalize the criteria necessary to designate a measure and the assessment process as evidence-based, thereby providing assistance in selecting the best instrument for an assessment

task. To date, two approaches have presented widely applicable options for categorizing evidence-based psychological instruments.

Hunsley and Mash (2008) provided an approach in which explicit criteria were used to evaluate the adequacy of psychological instruments, including measures used to assess frequently encountered difficulties (e.g., anxiety disorders, mood disorders, couple distress, personality disorder) within different populations (e.g., children, adolescents, adults, older adults, couples). These criteria were designed to determine whether measures where good enough for clinical or research use. Instead of focusing on the ideal criteria for a measure, the authors designed criteria that would indicate the minimum evidence needed to warrant the use of the measure. Ratings focus on nine criteria: norms, internal consistency, inter-rater reliability, test-retest reliability, content validity, construct validity, validity generalization, sensitivity to treatment change and clinical utility. What constitutes as adequate, good and excellent varied for each criteria but in general a rating of adequate indicates the minimal level of scientific evidence is met, good indicates solid scientific support and excellent indicates highly supporting evidence.

A different approach to classifying measures was taken in a recent task force of the American Psychological Association Society of Pediatric Psychology (Cohen, La Greca, Blount, Kazak, Holmbeck & Lemanek, 2008). The goal of the task force was to identify and systematically critique the psychosocial assessment instruments available to professionals working in the field of child health care, thereby providing a guide for identifying the instruments most appropriate for different purposes. In order to review measures in a systematic way, the task force developed specific criteria that closely resembled those used to identify evidence-based treatment. Depending on the available empirical support, each criterion could be evaluated as promising, approaching well-established, and well-established. Special attention was also placed

on the clinical and research utility of the measures, their applicability to various ethnic and linguistic minorities and whether they led to clear treatment implications.

Evidence-Based Assessment in Older Adults

Although research on evidence-based assessment is increasing, research on the assessment of older adults is much less developed (Ayers, Sorrell, Thorp & Wetherell, 2007). This is problematic as, with the growing number of older adults in the general population, there is also a concomitant rise in the number of older adults who require mental health services. The empirical literature on the mental health of older adults suggests that anxiety disorders have become a widespread problem in late life, with two-to-twelve-month prevalence estimates ranging from 1.2% to 15% in community samples of older adults (Bryant, Jackson & Ames, 2009; Wolitsky-Taylor, Castriotta, Lenze, Stanley & Craske, 2010) and from 1% to 28% in clinical samples of older adults (Bryant et al., 2009). Unfortunately, relatively little is know about the evidence based assessment of anxiety in older adults. As diagnosis and treatment selection is informed by assessment data, it is necessary to have measures that are appropriate for an older population, but the lack of research evidence for the psychometric quality of many of these instruments makes it challenging to choose an appropriate measure for use with older adults (Edelstein et al., 2008).

In a recent review of the literature, Therrien and Hunsley (2012a) found that there were a dozen anxiety measures commonly used in the assessment of older adults. After reviewing the research on these measures, Therrien and Hunsley concluded that most anxiety measures used with older adults were developed for a younger population and that, although there was empirical support for the use of some of these measures, the majority lacked sufficient evidence of their psychometric soundness with older adults. As having reliable scores is crucial, not only in ensuring that one is consistently measuring what one wants to measure but also in determining one's ability to detect the effects of interest, it is concerning that researchers commonly failed to report reliability coefficients for their own data.

In a subsequent study, Therrien and Hunsley (2012b) conducted a reliability generalization meta-analysis of these commonly used measures of anxiety in older adults in order to estimate their mean reliability score. Their meta-analysis focused on Cronbach's alpha, a measure of internal consistency that determines the degree of item agreement within a measure. The mean alpha coefficients found in this literature ranged from adequate to excellent making it possible to determine, on the basis of reliability, the strongest overall anxiety measures. Although meta-analytic studies such as these are ideal to gather information on the likely psychometric properties of a measure when used for a specific purpose and with specific populations, obtaining these estimates is extremely time-consuming and demanding to conduct. When faced with daily assessment tasks, researchers and clinicians need a faster way to judge whether a measure is appropriate and can be considered as evidence based. Classification schemes such as those developed by Hunsley and Mash (2008) and Cohen et al. (2008) could be a simple and appealing way to identify psychometrically sound measures without having to conduct more labour-intensive meta-analysis studies.

Purpose of the Present Study

Despite a common goal to apply standardized criteria for determining the extent of the evidence base supporting psychological measures, the classification approaches differ substantially in the criteria used to evaluate instruments. This raises significant questions regarding the utility and comparability of these approaches. In order to assess the comparability of the approaches, the reliability criteria of the Society for Pediatric Psychology (Cohen et al., 2008) and the Hunsley and Mash (2008) approaches were applied to the most commonly used measures of anxiety for older adults in order to determine if the resulting ratings are consistent with each another. Furthermore, the accuracy of the approaches were assessed by comparing the reliability criteria rating from both approaches for each measure with the mean reliability

estimates reported in the Therrien and Hunsley (2012b) reliability generalization meta-analysis. If the results obtained by the approaches yield similar results to the meta-analytic study (e.g., a measure that obtains a high mean reliability estimate in the meta-analysis also obtains a high rating by the classification approaches), researchers and clinicians may benefit from using the classification approaches and obtain similar results to more time-consuming meta-analysis.

Method

Selection of Articles

To prepare for the comparison of classification approaches, we used the databases PsycINFO and PubMed to identify articles that used at least one of the twelve most commonly used anxiety measures as identified by Therrien and Hunsley (2012a) in a previous systematic review. These were: the trait version of the State Trait Anxiety Inventory (STAI; Spielberger, Gorsush, Lushene, Vagg, & Jacobs, 1983), the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983), the Geriatric Mental State Examination (GMSE; Copeland et al., 1976), the Hamilton Anxiety Rating Scale (HARS; Hamilton, 1959), the Goldberg Anxiety and Depression Scale (GADS; Goldberg, Bridges, Duncan-Jones & Grayson, 1988), the General Health Questionnaire (GHQ, Goldberg, 1978), the Beck Anxiety Inventory (BAI; Beck, Epstein, Brown, & Steer, 1988), the Brief Symptom Inventory (BSI; Derogatis & Spencer, 1982), the Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, & Borkovec, 1990), the Symptom Checklist-90-Revised (SCL-90-R; Derogatis, 1994), the Geriatric Anxiety Inventory (GAI; Pachana, Byrne, Siddle, Kolowski, Harley, & Arnald, 2007) and the Worry Scale (WS; Wisocki, Handen, & Morse, 1986). Studies meeting the following criteria were selected: (a) an empirical study, (b) used at least one of the most commonly used anxiety measure with more than one participant, and (c) only included older adults aged 65 years and above in the sample. The search was restricted to articles published between January 1960 and August 2009.The first

search included the following terms: *anxiety*, *anxiety disorder, generalized anxiety disorder* AND *assessment*, *measurement*. The second search combined the following terms *anxiety*, *anxiety disorders*, *generalized anxiety disorder* AND *geriatric assessment*, *geriatric patients*, *geriatric psychiatry*, *geriatric*, *gerontology*. The final sample consisted of 142 studies.

Review of Measures

Three independent coders, one clinician and two graduate students, were instructed to focus on conducting reviews for ten of the twelve most commonly used anxiety measures with older adults. Two measures, (GMSE and HAMA), were not included in the present study as they are semi-structured interviews for which alpha coefficients do not apply. Coders were provided with each relevant articles and a coding manual for coding and classifying each of the measures. Given that the criteria established by Cohen et al. (2008) are more general (e.g., detailed information indicating good validity and reliability in at least one peer reviewed article) than the Hunsley and Mash (2008) criteria (e.g., preponderance of evidence indicating alpha values ≥ 0.90), coders were first asked to evaluate each of the twelve measures against the reliability criteria established by Cohen et al (2008). As the results will be compared to those obtained in a previous reliability generalization study that focused on internal consistency of the measures, coders were asked to concentrate on the Cronbach's alpha values reported in the articles.

The coding manual contained the three criteria established by Cohen et al. (2008) used to rate a measure's reported reliability values: the number of peer-reviewed articles conducted by different investigators, sufficient details about the measures to allow replication and detailed information indicating good reliability in at least one peer-reviewed article. Following the review of articles for each measure, coders were asked to determine whether each measure could be considered as promising, approaching well-established or well-established. See Table 1 for precise details about the criteria established by Cohen et al. (2008). As indicated in the coding

manual, the three coders then evaluated each measure against the reliability criteria established by Hunsley and Mash (2008) that focused on the preponderance of internal consistency evidence reported in the published research literature. Based on the extent of empirical support, coders were also asked to note whether each measure could be considered as adequate, good, excellent or has not having enough information. See Table 2 for precise details of the reliability criteria established by Hunsley and Mash (2008).

Results

Inter-Rater Reliability

An inter-rater reliability analysis using intraclass correlations (ICC) was performed to determinate the consistency among the three raters. In this study, a two-way random effects model was used as the raters were a random sample of a larger population of possible raters who coded targets (i.e., the anxiety measures) chosen from a larger pool of targets. We combined the multiple ratings, as this generally produces a more reliable measurement of inter-rater reliability (Nichols, 1998). Prior to this step, discrepancies in any of the ratings were resolved through discussions between raters.

The ICC can range between 0.0 and 1.0, approaching 1.0 when there is little variation between the scores given to each item by the raters (i.e., if all raters give the same or similar scores to each of the items). Put another way, ICC may be thought of as the ratio of variance explained by the independent variable divided by the total variance, where the total variance is the explained variance plus variance due to the raters and residual variance. The interpretation of kappa suggested by Landis and Koch (1977) can be used as a general guideline to evaluate the ICC. Therefore, to be defined as having satisfactory agreement, the ICC needs to have a value exceeding 0.60-0.70 (Gelfand & Hartman, 1975; Landis & Koch, 1977).

Coders were asked to determine whether each measure could be considered as promising, approaching well-established or well-established based on the Cohen et al. (2008) approach. The inter-rater reliability was satisfactory (ICC=0.75). For the Hunsley and Mash (2008) approach, coders were asked to determine whether each measure could be considered as adequate, good, excellent or as not having sufficient information. The inter-rater reliability was also found as satisfactory (ICC=0.79). The disagreements found in both approaches were mainly caused by a difficulty in knowing, in some articles, whether the reported alpha value was calculated for the study sample or if it was based on results of a previous study. An additional source of disagreement for the Cohen et al. (2008) approach stemmed from the fact that the criteria were not always interpreted in the same way by all coders. For example, coders needed to decide whether the reported alpha values were "moderate" or "good" but, without specific guidelines, coders sometimes differed in what they considered as moderate or good reliability. See Table 3 for the ICC values of each criterion.

Comparison of the Two Approaches

After resolving any discrepancies between raters, each measure received a rating based on the Cohen et al. (2008) approach (Well-established, Approaching Well-Established or Promising) and a rating based on the Hunsley and Mash (2008) approach (Excellent, Good or Adequate). We used Fisher's exact test to examine the relation between the ratings of the two approaches and consequently observe whether the probability of obtaining a certain rating was the same in both approaches. Fisher's exact test is a statistical significance test used in the analysis of contingency tables where sample sizes are small. Results suggest that the two approaches to classifying reliability in instruments do not yield the same results.

Comparison of the Two Approaches and Meta-Analytic Results

The mean alpha coefficient of each measure is available from a previously conducted meta-analysis (Therrien & Hunsley, 2012b). Thus, it is possible to compare this reliability information and the two classification systems. In order to do this, an exploratory analysis using a classification tree was conducted to determinate how the variables (Cohen et al. ratings, Hunsley and Mash ratings and the results from the meta-analysis) are associated. Classifications trees are a nonparametric procedure that do not require any assumptions about the distribution of the data and that allow the use of both categorical and continuous variables in a type of discriminant analysis with the end results being a graph that looks like a tree. The goal of the classification tree is to predict or explain responses on a categorical dependent variable (target) from their measurements on one or more independent variables (predictors). This allows us, for example, to predict whether a measure that obtained a rating of excellent (or Good or Adequate) on the Hunsley and Mash (2008) approach obtained a specific rating on the Cohen et al. (2008) approach or a certain mean alpha value on the meta-analysis.

The results of the classification tree revealed that there were no significant relations between the variables and that having a certain rating on one of the approaches did not allow us to predict the rating obtained on the other approach or the meta-analytically derived mean alpha value. However, these results are likely due, in part, to the small sample of measures as each measure was associated with only three values (i.e., a rating based on the Hunsley & Mash (2008) criteria, a rating based on the Cohen et al (2008) criteria and a meta-analytically derived mean reliability estimate). Moreover, a more descriptive look at the results (see Table 4) indicates that there was very little variability with the Cohen et al. (2008) approach, which also reduces the likelihood of detecting results in the classification tree. As there is little variability in the Cohen approach, most measures obtained the highest rating (e.g., Well-Established) and consequently did not seem to follow the mean reliability estimates derived from the meta-analytic study. On

the other hand, the Hunsley and Mash criteria (2008) led to more distinctions between measures (e.g., Excellent, Good and Adequate) that appeared to more closely follow the meta-analytic results.

Discussion

Recent efforts have been made in developing guidelines that operationalize the criteria necessary to designate a psychological instrument as evidence-based. Two approaches have presented widely applicable options for judging whether an instrument can be considered as evidence-based. Although both approaches have a common goal, they differ substantially in the criterion used to evaluate the instruments. This raises questions about the comparability and utility of the various efforts. The aims of this study were to (a) apply the reliability criteria of the Cohen et al. (2008) approach and the Hunsley and Mash (2008) approach to the most commonly used measures of anxiety with older adults in order to determine if the resulting ratings are consistent with one another and (b) compare the reliability criteria rating from both approaches for each measure with the mean reliability estimates reported in the Therrien and Hunsley (2012b) RG meta-analysis.

Of the 142 articles coded for this study, 69% did not provide any information at all regarding score reliability, 8% described reliability values as having been previously reported elsewhere, and 7% presented specific reliability coefficients from either the test manual or previous studies. Only 16% of published studies provided reliability coefficients from the data actually analyzed in the article. Despite repeated admonitions in the literature about the nature of reliability (e.g., Thompson & Vacha-Haase, 2000), these findings seem to reflect the tendency of researchers to view reliability values as a stable or invariant property of measures rather than as characteristics of the scores obtained from a specific sample. Even more surprising is that almost one-quarter of the coded articles provided the name of the measure and a reference without

giving any additional information about the measure such as the number of items or the population for whom it was created. Psychological measures are neither reliable nor unreliable in and of themselves, which is why researchers are advised to always describe measures and provide reliability coefficients of the scores for their data (Wilkinson & the APA Task Force on Statistical Inference, 1999).

Both approaches were created in order to give clinicians and researchers specific criteria that could help them make decisions regarding which measures are best suited for specific assessment tasks and, therefore, which measure can be considered as evidence-based. Results suggest that both approaches can be used with satisfactory inter-rater reliability. However, there is little consistency across both approaches suggesting that, not only do they use different criteria to evaluate the measures, but they also give different outcomes. As the criteria established by Cohen et al. (2008) are less demanding, eight of the ten measures (BAI, BSI, GAI, HADS, PSWQ, SCL-90-R, STAI and WS) were judged as well-established whereas, with the Hunsley & Mash (2008) criteria, only two of these measures (BAI and GAI) obtained a rating of Excellent and three of them (SCL-90-R, STAI and WS) obtained a rating of good. The difference between the two approaches seem to be caused by (a) the difference in the number of required studies reporting each measures ("two or more" vs "preponderance of published evidence") and (b) the precision of criteria for establishing psychometric adequacy. For example, alpha values reported for the BSI were: 0.74, 0.78 and 0.85. Using the Cohen et al. (2008) criteria, this measure obtained a rating of Well-Established, as at least one study reports a reliability coefficient that could be judged as "good." However, using the Hunsley and Mash (2008) criteria, this measure obtained a rating of adequate as the preponderance of evidence (two out of three studies) indicates alpha values of 0.70 to 0.79.

The two classification systems reviewed in this study attempt to apply standardized criteria for determining the extent and quality of the evidence base supporting psychological measures. However, despite a common goal, the two systems differ substantially in the criteria used to identify and evaluate the measures. The two systems clearly differed in the number of studies required to attain the best ratings, with the Cohen et al. (2008) system requiring at least one peer reviewed article indicating good reliability and the Hunsley and Mash (2008) system requiring a preponderance of evidence indicating adequate, good or excellent reliability. Considering the tendency to not report reliability values in research publications, as well as the common erroneous belief that measures themselves are reliable or unreliable, both of these approaches present with some specific challenges.

With respect to the Cohen et al. (2008) system, it is problematic that, for a measure to be considered as well-established, there needs to be only one peer-reviewed article providing evidence of good reliability. Considering that, as sample characteristics change it is possible and even likely that reliability estimates will also vary, it seems unlikely that only one example of good reliability can accurately represent the range of reliability values possible with a measure. Such an approach seems to reinforce the erroneous assumption that reliability is an attribute of the instrument. Because samples reveals a unique set of scores for a particular instrument, and because every samples reveals unique reliability coefficients, it cannot be assumed that every samples will yield equal even similar internal consistency reliability coefficients. Another challenge with this approach is having to subjectively rate "good" reliability with no further information on what can be considered as "good" reliability. Providing guidance in operationalizing "good" would undoubtedly greatly add to this approach and lead to enhanced agreement among those implementing this system. For example, in the present study, a

substantial portion of observed rater disagreements centered on whether a specific alpha value should be considered as "adequate" or "good".

On the other hand, the Hunsley and Mash (2008) approach requires that the preponderance of evidence should be considered in determining the classification of a measure. Although this approach solves the difficulty of subjectively having to rate levels of reliability by providing explicit guidelines, it assumes that a reviewer can examine all studies that used a specific measure in order to conclude whether there is a preponderance of reliability estimates. This is problematic because many of published studies do not include reports of a measure's reliability estimate based on the sample's data. For example, in the present review, only 16% of the coded articles provided the reliability values for their own data. This means that, although a measure might have been used frequently in the literature, it is possible that only one reliability estimate has been reported, thus making it impossible to judge the level of reliability available in the preponderance of research. The bottom line is that it is challenging, if not impossible, to operationalize "preponderance" when so many studies fail to report reliability values.

The results of the two classification approaches were compared to the results obtained in a previous reliability generalization meta-analytic study (Therrien & Hunsley, 2012b) as a way to determine whether they gave similar results than a more labor-intensive meta-analysis. It is important to note that the results obtained by the meta-analysis goes beyond what is reported in the literature as some reliability information used in the RG was obtained by contacting the authors through e-mail. The Cohen et al. (2008) system did not appear to distinguish the anxiety measures based on the mean reliability obtained in the RG as most measures obtained the highest rating (Well-Established) regardless of their mean reliability estimates. For example, the GAI obtained a mean reliability estimate of 0.92 in the RG study and the HADS a mean reliability estimate of 0.79 but, both measures obtained a rating of Well-Established. On the other hand, the

Hunsley and Mash (2008) approach is more conservative and uses more precise criteria that made it possible to make some distinctions that followed the results obtained in the RG. With some exceptions (e.g., the PSWQ had a mean reliability of 0.90 but a rating of adequate), the ratings obtained by the Hunsley and Mash (2008) approach were generally higher (e.g., Excellent or Good) in measures that also obtained a high mean reliability estimates in the RG study and a lower rating (e.g, Adequate) in measures that obtained a low mean reliability estimates in the RG study.

Limitations

A few limitations of the present study should be noted. First, as the data is based on a systematic review of the literature, the main limitation resides in the ability to identify and retrieve all of the studies that have used an anxiety measure with a sample of older adults. The main databases in the fields of health (PubMed) and mental health (PsycInfo) were searched, ensuring that the majority of relevant journals were considered. Unpublished studies were not included in the review and, as it is likely that unpublished studies contain reports of lower reliability coefficients than do published ones, the current sample of studies may yield an overestimate of reliability for some measures (Kieffer &MacDonald, 2011). Also, as very few studies reported the reliability coefficients for their own sample, our database could only include those studies that reported a reliability coefficient for the sample.

Although reliability is important when deciding whether a measure can be considered as evidence-based, a strong measure must also show evidence of other psychometric properties such as validity and, ideally, present evidence of clinical utility. Although both the Cohen et al. (2008) system and the Hunsley and Mash (2008) system consider ratings of validity, our study focused solely on the reliability criterion of both approaches in order to compare them to the findings of a previous reliability generalization meta-analytic study. Further studies will need to look at the

other criteria presented in both approach in order to determine if they lead to the identification of a similar set of evidence-based instruments. Additionally, other studies will need to assess the effectiveness of both approaches with instruments created for various purposes and age groups. If there is still little overlap among the instruments that can be identified as evidence-based, as was the case with the reliability criterion, it will be necessary to identify what constitutes as evidence-based instruments and what is the best method to evaluate whether or not a measure presents sufficient evidence to be judged as evidence-based.

Conclusion

Important steps have been taken in establishing specific criteria for evidence-based instruments and in addressing the scope of issues that must be addressed if EBA is to develop. Unfortunately, although research on EBA is increasing, research on the assessment of older adults is much less developed (Ayers et al., 2007). Consistent with the findings from the current study, a recent review of the literature suggested that reliability reporting practices in published articles have not improved over time and are not consistent with good scientific practices (Green, Chen, Helms & Henze, 2011). It is unlikely that reporting practices will change unless there is more emphasis placed on measurement properties in research training and the conduct of research across all disciplines working with older adults. It seems necessary that journal editors examine their policies regarding score reliability and require that reliability estimates be reported and that results be interpreted in light of such estimates. By doing so, it would encourage greater attention to measurement properties and appropriate scientific reporting practices.

There have been some efforts in recent years to develop approaches to operationalize evidence-based psychological instruments but, little is known about their comparability and validity. The present study examined the reliability criterion of two of these approaches (Cohen et al., 2008 and Hunsley & Mash, 2008) and found that not only do they use different criteria to

evaluate the measures, but they also yield different outcomes. The Cohen et al. (2008) approach is less demanding and consequently yielded few distinctions between measures regardless of the mean reliability estimates derived from meta-analytic research. On the other hand, the Hunsley and Mash (2008) approach is more precise and conservative, which led to more distinctions between measures that were in line with the meta-analytic results.

With the increasing proportion of older adult in the population and the frequency with which they experience anxiety, it is surprising that so little is known about the evidence-based assessment of anxiety in older adults. It seems that, due to a lack of assessment instruments created specifically for older adults, researchers and clinicians often need to rely on measures created for younger populations. However, as older adults experience anxiety differently than do younger adults, it is critical that the anxiety measures used with older adults be demonstrated to be scientifically sound. Researchers and clinicians working with older adults should first consider anxiety measures created specifically for older adults as they present with many advantages such as age-appropriate norms. If these measures do not appear appropriate for the sample at hand, then measures that were created for younger adults should be considered. Regardless of the chosen measure, researchers and clinicians need to determine whether it presents with good enough psychometric properties to warrant its use with an older population. The Cohen et al. (2008) approach to quickly judge whether a specific anxiety measure has the basic requirements necessary to potentially be considered as evidence-based in situations where a false positive decision is not too serious. On the other hand, the Hunsley and Mash (2008) approach can be used as an effective way to not only judge whether an anxiety is evidence-based when used with older adults but also to compare measures in order to decide which one is most appropriate for the sample at hand.

Although such approaches are an important step in the promotion of evidence-based practice and in the assessment of older adults, a major limitation resides in the fact that the criteria are developed for the identification of evidence-based instruments and not evidence-based assessments. One must consider that even when instruments have shown to provide some reliable and valid scores when used to assess anxiety in various sample of older adults, assessment is a complex decision-making task. In order to achieve a truly evidence-based approach to assessment, it will be necessary to not only assess the psychometric properties of such anxiety measures when used with older adults but also the accuracy and usefulness of the interpretations and conclusions drawn by clinicians using the instruments. As EBA is still at its early stage, it is likely that, as more information is gathered on which instruments can be considered as evidence-based, more attention will be paid on how it can contribute to EBA.

References

Ayers, C. R., Sorrell, J. T., Thorp, S. R. & Wetherell, J. L. (2007). Evidence-based psychological treatments for late-life anxiety. *Psychology and Aging, 22*, 8-17

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans,C. et al. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice–based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology, 69*, 184-196.

Barnes, L. L. B., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement, 62*, 603-618.

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: psychometric properties. *Journal of Consulting and Clinical Psychology, 56,* 893-897.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Belmont,CA: Chapman & Hall.

Bryant, C., Jackson, H., & Ames, D. (2009). Depression and anxiety in medically unwell older adults: Prevalence and short-term course. *International Psychogeriatrics, 21*, 754-763.

Cohen, L. L., La Greca, A. M., Blount, R. L., Kazak, A. E., Holmbeck, G. N., & Lemanek, K. L. (2008). Introduction to the special issue: Evidence-based assessment in pediatric psychology. *Journal of Pediatric Psychology, 33,* 911-915.

Copeland, J. R., Kelleher, M. J., Kellett, J. M., Gourlay, A. J., Gurland, B. J., Fleiss, & Sharpe, L. (1976). A semi-structured clinical interview for the assessment of diagnosis and mental state in the elderly: The Geriatric Mental State Schedule I: Development and reliability. *Psychological Medicine, 6*, 439-449.

Derogatis, L. R. (1994). *Administration, scoring and procedures manual*. Minneapolis, MN: National Computer System.

Derogatis, L. R., & Spencer, P. M. (1982). *The Brief Symptom Inventory (BSI): Administration, and procedures manual-I*. Baltimore, MD: Clinical Psychometric Research.

Edelstein, B. A., Woodhead, E. L., Segal, D. L., Heisel, M. J., Bower, E. H., Lowery, A. J., & Stoner, S.A. (2008). Older adult psychological assessment: Current instrument status and related considerations. *Clinical Gerontologist, 31(3)*, 1-35.

Gelfand, D. M. & Hartmann, D. P. (1975). *Child behavior analysis and therapy.* New York: Pergamon Press.

Goldberg D. et al. (1978). *Manual of the General Health Questionnaire*. Windsor, England: NFER Publishing.

Goldberg, D., Bridges, K., Duncan-Jones, P., & Grayson, D. (1988). Detecting anxiety and depression in general medical settings. *British Medical Journal, 297*, 897-899.

Green, C.E., Chen, C. E., Helms, J. E., & Henze, K. T. (2011). Recent reliability reporting practices in psychological assessment: Recognizing the people behind the data. *Psychological Assessment, 23*, 656-669.

Hamilton, M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology, 32*, 50-55.

Hatfield, D. R., & Ogles, B. M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology: Research and Practice, 35*, 485-491.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement, 61*, 404-420.

Hunsley, J., Lee, C. M., & Wood, J. (2003). Controversial and questionable assessment techniques. In S. O. Lilienfeld, J. M., Lorh, & S. J. Lynn (Eds.), *Science and pseudoscience in contemporary clinical psychology* (pp. 39-76). New York, NY: Guilford.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3,* 29-51.

Hunsley, J., & Mash, E. J. (2008). Developing criteria for evidence-based assessment : An introduction to assessments that works. In J. Hunsley & E. J., Mash (Eds.), *A guide to assessments that work* (pp.3-14). New York: Oxford University Press.

Kieffer, K. & MacDonald, G. (2011). Exploring factors that affect score reliability and variability in the Ways of Coping Questionnaire reliability coefficents : A Meta-analytic reliability generalization study. *Journal of Individual Differences, 32*, 26-38.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Mash, E. J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child and Adolescent Psychology, 34*, 362-379.

Meyer, T., Miller, M., Metzger, R., & Borkovec, T.D. (1990). Development and validation of the Penn State Worry Scale. *Behaviour Research and Therapy, 28*, 487-495.

Nichols, D. (1998). SPSS, Inc. *Choosing an Intraclass Correlation Coefficient*. Retrieved from http://www.utexas.edu/cc/faqs/stat/spss/spss4.html.

Norcross, J. C., Koocher, G. P., & Garofalo, A. (2006). Discredited psychological treatments and tests: A Delphi poll. *Professional Psychology: Research and Practice, 37*, 515-522.

Pachana, N. A., Byrne, G. J., Siddle, H., Koloski, N., Harley, E., & Arnold, E. (2007). Development and validation of the Geriatric Anxiety Inventory. *International Psychogeriatrics, 19*, 103-114.

Pande, A., Abdel-Aty, M., & Das, A. (2010). A classification tree based modeling approach for segment related crashes on multilane highways. *Journal of Safety Research, 41*, 391-397.

Shrout, P.E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 2*, 420-428.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P., & Jacobs, G. A. (1983). *Manual for the state-trait anxiety inventory STAI (Form Y): Self-evaluation questionnaire*. Palo Alto, CA: Consulting Psychologists Press.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (3rd ed.). New York: Oxford University Press.

Therrien, Z., & Hunsley, J. (2012b). *Assessment of Anxiety in Older Adults: A Reliability Generalization Meta-Analysis of Commonly Used Measures*. Manuscript submitted for publication.

Therrien, Z. & Hunsley, J. (2012). Assessment of anxiety in older adults: A systematic review of commonly used measures. *Aging and Mental Health, 16*, 1-16.

Thompson, B. & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195.

Wilkinson, L. & American Psychological Association of Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and Explanations. *American Psychologist, 54*, 594-604.

Wisocki, P. A., Handen, B., & Morse, C. K. (1986). The Worry Scale as a measure of anxiety among homebound and community active elderly. *The Behavior Therapist, 9,* 91-95.

Wolitsky-Taylor, K., Castriotta, N., Lenze E. J., Stanley, M. A., & Craske, M. G. (2010). Anxiety disorders in older adults: A comprehensive review. *Depression and Anxiety, 27*, 190-211.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression scale. *Acta Psychiatrica Scandinavica, 67*, 361-370.

Table 1

Criteria established by Cohen et al. (2008)

Well-Established assessment ▫Measure presented in at least two peer-reviewed articles by different investigators ▫Sufficient detail about the measure to allow critical evaluation and replication ▫Detailed information indicating good validity and reliability in at least one peer-reviewed article
Approaching Well-Established assessment ▫Measure presented in at least two peer-reviewed articles which can be by the same investigator ▫Sufficient detail about the measure to allow critical evaluation and replication ▫Validity and reliability presented in vague terms or moderate values
Promising assessment ▫Measure presented in at least one peer-reviewed article ▫Sufficient detail about the measure allow critical evaluation and replication ▫Validity and reliability presented in vague terms or moderate values

Table 2

Reliability criteria established by Hunsley & Mash (2008)

Internal consistency:
Adequate: Preponderance of evidence indicates alpha values of .70-.79.
Good: Preponderance of evidence indicates alpha values of .80-.89.
Excellent: Preponderance of evidence indicates alpha values ≥ .90.
Inter-rater reliability:
Adequate: Preponderance of evidence indicates kappa values of .60-.74; the preponderance of evidence indicates Pearson correlation or intraclass correlation values of .70-.79.
Good: Preponderance of evidence indicates kappa values of .75-.84; the preponderance of evidence indicates Pearson correlation or intraclass correlation values of .80-.89.
Excellent: Preponderance of evidence indicates kappa values of ≥ .85; the preponderance of evidence indicates Pearson correlation or intraclass correlation values of ≥ .90.
Test-retest reliability:
Adequate: Preponderance of evidence indicates test-retest correlations of at least .70 over a period of several days to several weeks.
Good: Preponderance of evidence indicates test-retest correlations of at least .70 over a period of several months.
Excellent: Preponderance of evidence indicates test-retest correlations of at least 0.70 over a period of a year or longer.

Table 3

Inter-rater Correlation Coefficients (ICC) for Both Classification Scheme's Coding

	Criteria	ICC value
Hunsley & Mash (2008) criteria	Reliability values (i.e., Coding alpha value for the study sample)	0.89
	Final rating (i.e., Excellent, Good, Adequate or Insufficient)	0.79
Cohen et al. (2008) criteria	Reliability values (i.e., No mention of reliability for the study sample, moderate reliability value or good reliability value)	0.93
	Sufficient details for replication (Coded as Yes or No).	0.76
	Final rating (i.e., Well-Established, Approaching Well-Established or Promising)	0.75

Table 4

Classification Schemes Coding and Meta-Analytically Derived Values

Measure	Meta-Analytically Derived values	Hunsley & Mash (2008) criteria	Cohen et al. (2008) criteria
GAI	0.92	Excellent	Well-Established
PSWQ	0.90	Adequate	Well-Established
SCL-90-R	0.90	Good	Well-Established
BAI	0.86	Excellent	Well-Established
WS	0.86	Good	Well-Established
GHQ	0.82	Insufficient	Promising
STAI	0.82	Good	Well-Established
BSI	0.81	Adequate	Well-Established
HADS	0.79	Adequate	Well-Established
GADS	0.74	Adequate	Approaching Well-Established

Conclusion

In the past decade, the health care system has witnessed a growing emphasis on the development, validation and dissemination of evidence-based practices, suggesting that when considering a health care service for a patient, professionals must integrate information from systematically collected data, clinical expertise and patient preferences (Institute of Medicine, 2001). Evidence-based practice is now a central component of most health care systems and has been promoted through reports of various professional organizations (e.g., Institute of Medicine, 2001; American Psychological Association Presidential Task Force on Evidence Based Practice, 2006). With increasing demands that clinicians use brief and effective treatments, the field of psychology has placed a great deal of attention on the development of evidence-based treatments (EBT). Today, many of these treatments are taught in psychology classes, presented in workshops and described in various books. Given the advancement made in the area of EBT, it is surprising that, until recently, limited attention has been placed on identifying evidence-based approaches to assessment.

Psychological testing and assessment has been a central integral component in the development of clinical psychology and it remains one that is evident in contemporary psychology. Many clinical psychologists see their expertise in assessment as a unique feature of their profession that differentiates them from other health care disciplines (Hunsley & Mash, 2007). In a recent survey of 549 clinical psychologists randomly selected from the American Psychological Association (APA) Division of Clinical Psychology, the majority of respondents reported being routinely involved in some diagnosis or assessment activity (Norcross & Karpiak, 2012). However, the successful history of measurement, test development, and assessment within psychology has left the field with a confidence regarding the value of existing psychological assessment methods (Hunsley & Mash, 2010). Consequently, there seems to be a tendency to uncritically adopt a measure and assume that it summarizes what it is suppose to measure. There

also appears to be a belief that once a measure or a test has been published and some evidence of its psychometric properties are available, it must be reliable for any evaluation purpose.

Despite important advancement in the field of test development, Piotrowski (1999) noted that the ranking of the most commonly used instruments by clinical psychologists has barely changed over the and, not surprisingly, multiple studies have since found that many commonly used clinical assessment methods and instruments lack evidence to support their usefulness (Norcross, Koocher, & Garofalo, 2006). Fortunately, over the past decade, we have seen a modest but noticeable shift in the assessment literature that led to the development of evidence-based assessment (EBA). Although research in the field of EBA is increasing, it seems that, as is often the case, much less attention is being paid to the older adult population and, consequently, research on the assessment of older adults is much less developed (Ayers et al., 2007). With increasing proportions of older adults seeking treatment, mental health professionals need to develop greater awareness and understanding of psychiatric disorders present in late life in order to deliver effective treatments. Anxiety has become a widespread problem in late life but remains under-detected and under-reported (Byrne & Pachana, 2011). The first step in initiating appropriate treatments for older adults with anxiety is to have appropriate assessment instruments for this specific population. As older adults experience anxious symptoms in a different way than younger adults do and as there are many measures with older adults, it is critical that the anxiety measures used with older adults be demonstrated to be scientifically sound.

Anxiety Measures Most Commonly Used with Older Adults

Results of our systematic review (Study 1, Therrien & Hunsley, 2012a) suggest that with more than 90 measures used to evaluate anxiety in older adults, there does not seem to be a consensus on which measure is more appropriate for the geriatric population. Of those, ten measures appeared to be more frequently used with older adults: BAI, BSI, GADS, GHQ, GMSE, HADS, HARS, PSWQ, SCL-90-R and STAI. Surprisingly, only one of the most commonly used anxiety measures, the Geriatric Mental State Examination (GMSE), was developed specifically for older adults. This suggests that the lack of anxiety measures created

and validated for older adults often forces researchers and clinicians to rely on anxiety measures created for younger adults when assessing older adults. However, because many differences exist between younger and older adults, it is unlikely that a single measure can adequately assess anxiety across the entire lifespan and authors rarely mentioned whether measures were appropriate for older adults.

Most of the assessment measures reviewed in Study 1 lack sufficient evidence for their psychometric soundness when used with older adults and several critical considerations limit their use with geriatric populations. First, although each reviewed measure showed adequate internal consistency, the existing data often comes from a single published study and mostly from samples of older psychiatric outpatients. Both replication and extension of previous findings are necessary in order to determine if the measures are good general screening tools for anxiety in older adults. Additionally, most measures lacked evidence of both discriminant and concurrent validity as well as test-retest reliability. Taken together, only three of the most commonly used measures showed sufficient psychometric properties to justify their use with older adults. The BAI, a measure of general anxiety, and the PSWQ, a measure of worry, have both demonstrated high internal consistency as well as evidence of validity, suggesting that they might be good choices to consider when selecting an anxiety measure for a geriatric population. These results are in line with the findings of the reliability generalization study (i.e., Study 2, Therrien & Hunsley, 2012b) in which the BAI and the PSWQ showed high reliability estimates. This suggests that, based on their overall level of reliability and previous psychometric evidence, both researchers and clinicians assessing anxiety in older adults should consider these measures as likely to be the best currently available. The GMSE, a semi-structured interview that was created for older adults and that includes anxiety and phobia subscales, is the only commonly used measure that provides norms and cutoff scores validated for older adults. That, combined with its appropriate psychometric properties, suggests it is a good option for assessing anxiety in older adults. However, the GMSE has not been used in research published in the last decade, which may suggest that more recent measures have since been developed and are possibly more

appropriate for the current cohort of older adults.

Six of the reviewed measures (BAI, BSI, GHQ, GADS, HARS and SCL-90-R) were heavily weighted with somatic symptoms, which make it difficult to distinguish between anxiety symptoms, symptoms of other health problems or normal aging among the geriatric population. This can be problematic, as not only are older adults more likely to experience somatic symptoms when anxious, but they are also more likely to have coexisting physical conditions that may produce anxiety-like symptoms. Another issue that must be taken into consideration when evaluating the presence of anxiety in older adults is the frequent co-existence of anxiety and depression in later life. As these comorbid conditions can increase the complexity of anxiety assessment and diagnosis, it is important that the measures differentiate, as much as possible, anxious and depressive symptoms in older adults. However, none of the reviewed measures showed adequate evidence of discriminant validity with respect to mood disorders. High correlations between measures of depression and anxiety can be understood as resulting from the over-inclusion of non-specific symptoms that are present in both conditions. Additionally, most measures do not present age appropriate norms or clinically relevant cut-off scores, which greatly limit their use with older adults. Considering the limited research on the psychometric properties and the shortcomings of the current approaches to assessing anxiety in older adults, clinicians and researchers must be cautious and carefully consider the strengths and weaknesses of different anxiety measures before deciding which one to use with an older population. Using measures for which psychometric properties are still in question may lead to an invalid assessment and diagnosis of anxiety in older adults.

It must be underlined that, in the past decade, considerable effort has been made to develop anxiety measures specifically suited for older adults. Based on the Study 1 systematic review, two measures that were specifically created for older adults were sufficiently used to allow for some replication of psychometric findings: the Worry Scale (WS) and the Geriatric Anxiety Inventory (GAI). Both measures showed sufficient psychometric evidence to warrant their use in

assessing anxiety in older adults. In many instances, these instruments might be preferred over more commonly used measures that were not developed for the geriatric population and whose psychometric properties remained largely unknown with regard to older adults. Measures created specifically for older adults, such as the GAI and the WS, have many advantages over measures developed for younger adults including superior psychometric properties, norms established for an older population and increased acceptability among older adults (Edelstein & Segal, 2010). Additionally, they can help in addressing specific late life issues such as how developmental and medical factors influence anxiety symptoms.

Considering those advantages, clinicians and researchers looking for a general measure of anxiety to use with older adults should consider using the GAI. The GAI is not a diagnostic tool but rather a simple instrument that was designed to measure symptom severity. As such, it can be used as a brief tool of the experience of anxiety across the range of anxiety disorders (Byrne & Pachana, 2011). In the original article describing the development of the GAI (Pachana et al., 2007), it showed good psychometric properties in community-dwelling older adults and older patients attending a psychogeriatric service. On the other hand, clinicians and researchers wanting to assess the presence of worry in older adults should consider the WS. The WS was first developed by Wisocki et al. (1986) to assess worries particularly relevant to older adults (i.e., the areas of financial, health and social concerns). The scale has been revised through many studies and contains three new scales: world issues, personal concerns and family issues. Initial evidence suggested good psychometric properties for the total score as well as each subscales for older adults with generalized anxiety disorder and those without diagnosable psychiatric disorders (Stanley, Beck & Zebb, 1996).

Evidence-based Assessment Initiatives and Criteria

It is important to keep in mind that psychometric characteristics are not properties of an

instrument but rather properties of an instrument when used for a specific purpose and for a specific population. Consequently, there is no agreement as to the minimum psychometric value a measure must meet in order to be considered as scientifically sound. This is of little help to researchers and clinicians who need to judge whether a measure is appropriate when faced with daily assessment tasks. Fortunately recent efforts have been made in developing guidelines to operationalize the criteria necessary to designate a measure as evidence-based (e.g., Bickman et al., 1999; Cohen et al., 2008; Hunsley & Mash, 2008b). However, despite a common goal, these initiatives differed in the criteria they used to identify which measure can be considered as evidence-based and the manner in which the criteria were applied to the instruments.

Although the differences in the various approaches are likely due to the fact that the attention paid to EBA is quite recent, it does raise questions regarding the comparability and the validity of the approaches. In order to assess the comparability of the approaches, two of the three approaches (Cohen et al., 2008 and Hunsley & Mash, 2008b) were examined in Study 3 by applying them to the most commonly used measures of anxiety in older adults in order to determine whether the ratings are consistent across approaches. Results suggested that both classification systems can be used with satisfactory inter-rater reliability but that there is little consistency across approaches, suggesting that not only do they use different criteria to evaluate measures but also that they also yield different outcomes. The differences observed between both approaches seems to be explained by the number of required studies supporting each measure ("preponderance of evidence" vs "two studies") and the precision of criteria for establishing psychometric adequacy ("alpha values above 0.90" vs "good reliability"). Subsequent work will therefore need to examine the other criteria of the approaches and identify what constitutes evidence-based instruments (EBI) and what is the best method to determine which instrument can be considered as evidence-based.

Furthermore, the accuracy of the approaches was assessed by comparing the reliability criteria ratings from both approaches for each measure with the mean reliability estimates reported the reliability generalization meta-analysis (Study 2, Therrien & Hunsley, 2012b). Results suggested that the Cohen et al. (2008) approach does not appear to distinguish the anxiety measures based on the mean reliability estimates obtained in the RG, as most measures obtained the highest rating (Well-Established) regardless of their mean reliability estimates. This approach might be better used as a way to evaluate whether a measure has the basic requirements necessary to potentially be considered as evidence-based. The Hunsley and Mash (2008a) approach is more precise and conservative which made it possible to make some distinctions that followed the results obtained in the RG. Researchers looking to compare various measures and decide which is more appropriate for the assessment task should favor this approach.

Reliability Reporting Practices in Reviewed Articles

The reporting and use of reliability estimates in peer-reviewed articles has often been criticized (e.g., Helms, Henze, Sass, & Mifsud, 2006; Vacha Haase, Kogan, & Thompson, 2000). Cronbach's alpha, a measure of internal consistency is the most frequently used, but also the most commonly misused, reliability estimate (Hogan, Benjamin, & Brezinski, 2000). Typically, researchers have reported internal consistency as a property of a specific sample's response to a measure administered under specific conditions (Thompson & Vacha-Haase, 2000). It is not uncommon, even among experienced researchers, to describe the "reliability of a test" or mention that "the test is reliable". For example, Wilkinson and the American Psychology Association Task Force on Statistical Inference (1999) highlight that it is "important to remember that a test is not reliable or unreliable" (p.596), but they also mistakenly mention in the next paragraph "besides showing that an instrument is reliable." This seems to have led to a widespread

erroneous belief that a measure is either reliable or unreliable and, consequently, many researchers do not report coefficients of internal consistency for their own data.

Many authors have attempted to address the improper reporting and use of reliability coefficients and more particularly of Cronbach's alpha in the research literature. For example, Wilson (1980) reviewed articles published in the American Educational Research Journal (AERJ) between 1969 and 1978, and concluded that reliability was unreported in much published research and that it was inexcusable. He advised researchers to report reliability coefficients for their own set of data and argued that editors and reviewers should return papers that fail to establish psychometric properties. Almost 20 years later, Vacha-Haase, Ness, Nilsson, and Reetz (1999) reviewed three journals and found that only 36% of the quantitative articles provided reliability coefficients for the data being analyzed. More recently, Green, Chen, Helms, and Henze (2011) reviewed past and current reliability reporting practices in a sample of Psychological Assessment articles published across three decades and found that reliability practices have not improved over time and generally are not consistent with good practices. Other recent RG analyses have found similar rates to these previous values (Graham, Diebels, & Barnow, 2011; Lopez-Pina, Sanchez-Meca, & Rosa-Alcazar, 2009; Wheeler, Vassar, Worley, & Barnes, 2011).

In the RG study (Study 2, Therrien & Hunsley, 2012b), an astounding 69% of the reviewed studies made no mention of reliability. An additional 15% of the studies used reliability induction and mentioned score reliability but merely provided previously reported values as if the applied to their data by referencing the manual or a previous study. Using reliability induction is problematic as the magnitude of reliability estimates are related to various sample characteristics such as sex and ethnicity (Dawis, 1987). Because every sample has a unique characteristic that will lead to different set of scores on a measure and because every set of scores has a unique

internal consistency reliability coefficient, it cannot be assumed that the reliability statistics presented in test manuals will be equal or even similar to the sample at hand. In addition, several researchers (e.g., Caruso, 2000) have found that the score reliability estimates vary significantly on different administration of the same instrument. In our RG study, most of the commonly used measures for older adults were created for younger adults and inducting reliability from the test manuals is not only insufficient but also inappropriate. Furthermore, the reliability estimates of the reviewed measures were significantly different between samples. For example, the alpha value of the STAI ranged from 0.32 to 0.92, suggesting that a measure cannot be considered as reliable or unreliable and that it is essential to report reliability estimates for of some of the data at hand.

The tendency to not report reliability coefficients for one's own data seems to have prevailed despite the guidelines published by various professional organizations. In their effort to improve the quality of psychological assessment, the American Psychological Association (APA) struck the Task Force on Statistical Inference in order to provide some recommendations on the use of various statistical methods (Wilkinson & APA TFSI, 1999). When it comes to reliability, the APA Task Force advised authors to "provide reliability coefficients of the score for the data being analyzed even when the focus of their research is not psychometric (Wilkinson & APA TFSI, 1999, p.596). Additionally, the APA Ethical Principles of Psychologists and Code of Conduct (2002) highlights the importance of using assessment instruments that have shown evidence of reliability and validity with members of the population being tested and that, when this information is unavailable, the strengths and limitations of test results and interpretation must be described.

Guidelines for sound psychometric practices of reporting and interpreting reliability are also presented in the Standards for Educational and Psychological Testing (American Educational

Research Association et al., 1999) and the Code of Fair Testing in Education (Joint Committee on Testing Practices, 2004). The Standards for Educational and Psychological Testing (AERA et al., 1999) established generic standards to follow when developing and using psychological instruments. When it comes to reliability, AERA et al., (1999) states that " the critical information on reliability includes…a description of the examinee population for whom the foregoing data apply as the data may accurately reflect what is true of one population but misinterpret what is true of another" (p.27). Along the same lines, the Joint Committee on Testing Practices (2004) goes a little further and recommends that professionals not only provide evidence on the performance of test takers of diverse subgroups but that they also make sure that they ensure that the differences in performance are related to the skills being assessed.

Although various guidelines have highlighted the potential harm in assuming that the reliability evidence of one sample can generalize to other sample, they do not seem to have had the desired impact, as researchers do not seem to have changed their practice when reporting reliability estimates. This might be due to the fact that most guidelines simply state the importance of reliability but give no concrete strategies on reporting, analyzing and interpreting reliability. Researchers may well intend to follow appropriate practices for reporting and interpreting reliability estimates but have no idea of what such practices might be. To assist with this, Helms, Henze, Sass, and Mifsud (2006) defined seven broad categories (see Table 1) of good practices for reporting (e.g., calculate and report appropriate reliability coefficient for each measure used), analyzing (e.g., calculate and report confidence intervals for all reliability coefficients), interpreting (e.g., conduct RG analyses to aid in comparing coefficients across studies) and using (e.g., specify the intended use of the reliability coefficients *a priori*) reliability that is consistent with the Standards for Educational and Psychological Testing (AERA et al., 1999). They also described pragmatic strategies for implementing the good practice of reliability

in quantitative studies. These recommendations are sound and should be used as a guide whenever one is reporting, analyzing, interpreting and using reliability data. Hopefully, implementing good practices would address the erroneous beliefs surrounding reliability and increase confidence in not only the measures used in psychology but also in the results of studies obtained using such measures.

Based on the various guidelines and the recent effort of Helms et al. (2006) I would like to highlight a few recommendations to improve the practice of reporting reliability in the current literature. First off, consistent with the recommendations of the APA Task Force (Wilkinson & TSFI, 1999) and of others (e.g. Helms et al., 2006; AERA et al., 1999), I strongly suggest that researchers need to report reliability coefficients for their own data. In order to calculate coefficients of internal consistency researchers simply need access to the raw data obtained on the measure. There is therefore no excuse for not computing and reporting an internal consistency estimate of reliability when such information is available. In the unlikely event that the researcher does not have access to the item-level data, researchers should minimally report reliability estimates reported by other studies using similar samples and compare their sample with the inducted group (Vacha-Haase et al., 2000). Reporting internal consistency not only provides important information about the data in the study but also allows researchers to conduct RG meta-analytic study.

Additionally, as mentioned by Onwuegbuzie and Daniel (2002), I would like to highlight the importance of using confidence intervals when reporting reliability and interpreting sub-sample reliabilities in group comparison studies. Considering that reliability coefficients represent estimates that are subject to errors, the use of confidence intervals is appropriate (Onwuegbuzie & Daniel, 2002). More precisely, reliability estimates represent a lower bound for the theoretical reliability coefficients (Crocker & Algina, 1986), suggesting that, in addition to

reporting reliability for their own data, researchers should provide the upper confidence limits of the reliability coefficients. Furthermore, it is important for researchers to not only report the reliability coefficients and confidence intervals for the full sample at hand but also for all relevant subgroups. For example, researchers comparing a sample of younger adults and a sample of older adults should report the reliability coefficient for both age groups. As mentioned by Onwuegbuzie and Daniel (2000), for subgroups to have equal reliability estimates, both the variance of the total instruments scores and the sum of the individual item variances must be equal for all subgrougs. However, score reliability is likely to vary across subgroups in a sample and reporting only the reliability for the full sample might be inappropriate.

One can speculate that reliability estimates might not be reported in some studies as it presents a low value and researchers worry that the current data are useless. However, in such cases, one must not discard the data but rather examine the results in order to find out why low reliability estimates were obtained for the current sample if high reliability estimates were obtained in other studies. This should involve the examination of measures of sample homogeneity and item response (Onwuegbuzie & Daniel, 2002). By deciding to simply not report low reliability estimates, researchers risk presenting biased results, as internal consistency has an impact on statistical power (Onwuegbuzie & Daniel, 2000), with low internal reliability reducing statistical power. As advanced by Onwuegbuzie and Daniel (2002), when the null hypothesis is not rejected and one or more measures has scores with a low internal consistency, it is not clear whether the non-significant result suggests feasibility of the null hypothesis or if it is the result of a statistical artifact.

When it comes to the assessment of anxiety in older adults, there should be efforts to the core constructs relevant to the presentation of anxiety in older adults. As mentioned in numerous anxiety studies, many older adults present significant anxiety symptoms that do not meet criteria

for an anxiety disorder, suggesting that older adults do not experience anxiety in the same way as younger adults do (Kogan et al., 2000). It is therefore essential to identify the core domains of anxiety in older adults as it is not clear whether many current anxiety assessment instruments address the unique characteristics of the experience and presentation of older adults. This would not only help in developing but also validating evidence based assessment instruments and protocols that could serve as gold standards against which additional measures could be compared. Once core domains and instruments are identified, this information should be made available to practicing researchers and clinicians through the publication of expert consensus guidelines and also to those early in professional training in various health related graduate training programs.

Other Important Factors in Building an Evidence-Based Approach to Assessment

Although recommendations made in the present dissertation are primarily based on reliability, it is important to note that, at a minimum, for a measure to have any clinical use, information regarding both reliability and validity indices must be available. As mentioned by Hunsley and Meyer (2003), replicated evidence for a measure's concurrent, predictive and discriminant validity is necessary for a measure to be considered as potentially evidence-based. Just as the case with reliability, validation is a context-specific concept and therefore a measure must have validity evidence for each population it is used with (Hunsley & Mash, 2008a). In Study 1 (Therrien & Hunsley, 2012a) systematic review, the majority of the most commonly used anxiety measures with older adults lacked evidence of both discriminant and concurrent validity, suggesting a pressing need for the validation of measures that were created for young adults and that are frequently used with older adults. Having discussed the importance of reliability and validity in evidence-based assessment, the following sections will highlight some of the psychometric properties that need to be addressed in order for psychological assessment to be

truly evidence-based. These include the importance of age-appropriate norms, the role of incremental validity and the need for evidence on clinical utility.

Norms

Researchers must keep in mind that even when instruments are generally reliable and valid, an evidence-based measure should also possess appropriate norm-referenced interpretation and replicated supporting evidence for accuracy of cut scores. Norms or specific criterion-related cut-offs scores are necessary if one wants to interpret the results of a measure in a meaningful way (AERA et al., 1999). For example, knowing that an older adult obtained a score of 20 on an anxiety measures tells nothing unless we know the range of scores obtained by other older adults and, in case of clinical instruments, the cutoff score value that indicates whether a score falls in the "clinical" or the "normal" range. As mentioned by Achenbach (2001), because of the diverse range of clients for which most instruments are used, it is necessary to have nationally representative norms that are sensitive to gender and age influences. As few anxiety measures are originally created for older adults, it is unlikely that age-appropriate norms are available for interpreting a score, thus making the interpretation of the results quite challenging. In Study 1 (Therrien & Hunsley, 2012a) systematic review, only one measure (GMSE) had been normed for older adults and provided cutoff scores validated for this population. This suggests that most anxiety measures used with older adults are severely limited in their clinical value.

Incremental Validity

Researchers must keep in mind that even when measures are reliable, valid and possess appropriate norms, the way in which they are used in clinical practice may not be reliable and valid. Consequently, scholars of clinical psychology have discussed the concept of incremental validity for over 50 years. The definition of incremental validity differs across various researchers (e.g., Barnett, Lentz & Macmann, 2000; Cronbach & Gleser, 1957; Elliott,

O'Donohue, & Nickerson, 1993; Mischel, 1968) in the degree to which they incorporate various concepts such as cost effectiveness. However, the core element across definitions remains the idea of relative predictive efficacy: does a measure add to the assessment process in a way that goes beyond what other methods can? Incremental validity can therefore be defined as the degree to which a measure either explains or predicts what is being assessed (e.g., anxiety) relative to other measures. For example, does using a measure such as the Beck Anxiety Inventory provide better results during a 50- minute intake interview than just conducting the interview alone?

Incremental validity rises primarily from financial, practical and ethical concerns. Using a long and expensive assessment battery to either make a diagnosis or plan a treatment makes no sense if a briefer and less expensive method would give the same results. As highlighted by the Canadian Psychological Association (CPA) and the APA, psychological services such as assessment needs to be implemented with the client's best interest, and using lengthy and costly instruments with no evidence of incremental validity is unlikely to be in the best interest of the client. In spite of the attention paid to incremental validity in the last decades, it is rarely evaluated in new clinical assessment measures. Haynes and Lench (2003) reviewed the 298 manuscripts submitted to Psychological Assessment between 1998 and 2002 that reported on the development and validation of a new instrument scale and found that only 26 of the manuscripts addressed the incremental validity of the new instruments. Given the various measures used to assess anxiety in older adults, surprisingly little information is available on the extent to which data from one instrument improves upon what can be obtained from other instruments. Consequently, there is little evidence in the literature to aid in determining which measures to use to assess anxiety in older adults, how to combine them and what to do if conflicting results are obtained with some measures. More research is necessary before clinical psychologists will be

able to make scientifically informed decisions about which anxiety measure to use when assessing an older population and which might be redundant and ineffective.

Clinical Utility

Clinical utility is an increasingly used concept in the evaluation of both diagnostic systems (e.g., First et al., 2004; Kendell & Jablensky, 2003) and psychological assessment instruments (e.g., Hunsley & Bailey, 1999; McFall, 2005; Yates & Taub, 2003). With respect to psychological instruments, clinical utility refers to whether, when taking into account practical considerations (e.g., cost, ease of administration), there is evidence that the use of test data improve decision making or demonstrates clinical benefit such as better treatment outcome or greater patient satisfaction with services (Hunsley & Mash, 2008a). To date, much of the literature in the field of psychological assessment has focused on psychometric properties, whereas little attention has been paid to clinical utility (McGrath, 2001). Consequently, there is little evidence that bears on the question of the extent to which EBIs or EBA have a direct impact on the improved provision and outcome of clinical services. At present, therefore, for the majority of psychological instruments, a determination of clinical utility must often be made on the basis of likely clinical value, rather than on empirical evidence.

As previously mentioned, most of the assessment measures reviewed in Study 1 were created for younger adults and lacked sufficient evidence for their psychometric soundness when used with older adults. Not only is there a need for both replication and extension of previous findings that examine the reliability and validity of most measures, but it is also necessary to examine their clinical utility and develops appropriate norms. If research demonstrates that the early accurate identification of anxiety in older adults with one of the anxiety measures led to better treatment outcomes and less premature termination of services, then a very strong case could be made for the use of such assessment with older patients suspected of having an anxiety

disorder. With the aging population, there is a growing number of anxiety instruments such as the GAI and the WS designed for use with older adults. Although these measures present with evidence of psychometric soundness, there is limited information that considers the extent to which inclusion of these measures consistently improves upon clinical decision making and the outcome services.

Although psychometric theories view clinical utility as one of the many properties of a measure, other approaches to assessment such as clinimetrics (Feinstein, 1987) and communimetrics (Lyons, 2009) emphasize utility as the most important characteristic of an instrument. Whereas psychometric measures are mainly designed for research purposes, clinimetrics and communimetrics assessment tools are designed to be congruent with the clinical process. Assessment tools created by both these approaches are therefore designed so they can operate at the item level. In other words, because these approaches are action oriented, the items that are included in the measures have meaningful and direct links to what happens next in the services. The individual items are selected to guide decision making, as they indicate what level of service effort is required (i.e., 0 = No evidence, no need for action, 1 = Watching waiting or monitoring only required and 2 = Action is needed). Consequently, are designed to be accessible to service providers, consumers and policy makers.

For example, the Child and Adolescent Needs and Strengths for Children and Youth with Mental Health Challenges (CANS-MH) is a multi-purpose tool created for children's services in order to support decision making and to facilitate the linkage between the assessment process and the creation of individualized services plans. Developed from a communimetrics perspective, each item suggest different pathways to services. More specifically, each item is anchored with a definition describing the action levels that needs to be taken. For example, the level of suicidality of the child/adolescent is rated by the clinician on a four point scale (0 = Child has no evidence or

history of suicidal behaviours, 1= History of suicidal behaviours but no suicidal behaviour during the past 30 days, 2 = Recent (last 30 days) but not acute (today) suicidal ideation and 3 = Current suicidal ideation and intent in the past 24 hours) with ratings of 2 or 3 indicating the need for a safety plan.

Attention to clinical utility in psychometric theories has grown in recent years but, unfortunately, there is still little evidence of the extent to which EBIs and EBA have clinical utility. A truly evidence-based approach to clinical assessment requires psychometric evidence of the soundness of instruments and strategies but also, as highlighted by communimetrics, data on the fundamental question of whether or not the assessment process makes a difference with respect to the accuracy, outcome or efficiency of clinical activities. Psychometric measurement would benefit from considering clinical utility as an essential component of assessment and from evaluating the extent to which assessment instruments are acceptable to clients, enhance the quality and outcome of clinical services and are worth the cost associated with their use. Regardless of the purpose of the assessment, the central focus within EBA on the clinical application of assessment strategies makes it clear that there is a need for research on clinical utility (Hunsley & Mash, 2007).

The Relation Between Evidenced-Based Assessment and Evidence-Based Instruments

Even with measures showing strong psychometric properties, it is important to remember that the assessment process is a complex decision-making task in which information may be gathered in multiple settings, by multiple informants and at multiple times. In order to achieve a truly evidence-based approach to assessment, it will be necessary to not only assess the properties of individual assessment instruments but also the accuracy and usefulness of this complex decision-making task in light of potential errors of interpretation. However, due to the early stage of development of EBA, reviews of EBAs appear to be limited to the identification of evidence-

based instruments (EBI). It is impossible to develop an evidence-based approach to the assessment of anxiety in older adults without having done rigorous psychometric evaluation of the individual assessment instruments.

The main challenge in the dissemination of EBA in the field of anxiety in older adults resides in the fact that we do not have an agreed-upon list of EBIs. The measures that have shown strong psychometric properties in the current dissertation (BAI, GAI, GMSE, PSWQ and WS) would be good starting point for such work. Subsequent steps will need to ensure that the instruments are appropriately administered, scored and interpreted in accordance to the scientific literature. Although, the creation of an evidence-based protocol is the desired outcome in the field of anxiety in older adults, the best advice at this time might be to make sure that assessments are conducted with EBI and in a way that is guided by research evidence. Over time, it is likely that increased attention will be paid to the question of how data from EBIs can contribute to the field of EBA.

Obstacles for Identifying Evidence-based Assessment Procedures

Many scientific and logistic challenges will need to be addressed in the identification and dissemination of evidence-based assessment procedures. For example, whereas the APA Division 12 Task Force identified 6 well established and 13 probably efficacious treatments, Antony, Orsillo, and Roemer (2001) identified more than 200 anxiety-related assessment instruments that had considerable empirical support. In Study 1 (Therrien & Hunsley, 2012a) there were over 90 measures that had been used in the literature to assess the presence of anxiety in older adults. Additionally, given the frequency with which new assessment tools are developed, it will be difficult to evaluate the psychometric evidence of each one and judge whether they possess sufficient empirical support to consider them evidence-based. Any list of evidence-based

assessment instruments, as well as evidence-based assessment procedures, will need to be continually updated.

The purpose of treatment is relatively simple and straightforward. A treatment will generally be considered as effective when it alleviates symptoms and reduces distress. On the other hand, assessment has many different purposes, making it much more difficult to identify evidence-based assessment procedures than evidence-based treatments. For example, the role of assessment can be to establish a diagnosis, to select an intervention, to measure treatment outcome or to decide whether to include or exclude participants from a research study (Antony, 2002). Therefore, before determining whether an assessment instrument can be considered as evidence-based, one must determine for whom and for what purpose is the instrument evidence-based. For example, a scale might be valid for the purpose of assessing the severity of anxiety symptoms in older adults but it might not be helpful in selecting an appropriate treatment. In establishing evidence-based assessment procedures, it will be important to understand the context for which the instrument is evidence-based.

One of the main elements of EBA is that one should only use instruments that show good psychometric properties. However, similar to the debates regarding how many supporting studies are needed to conclude that a treatment is evidence-based, the field of assessment still struggles in defining what are good psychometric properties. At this time, there is no specific agreed upon cutoffs of reliability and validity that can be used when judging whether an instrument is evidence-based. It will be important to establish some consensus on the psychometric qualities necessary for an instrument to merit its use in clinical services. Additionally, it is possible, as is the case with evidence-based treatment, that professionals might be hesitant in giving up using assessment methods that they frequently use but that are not evidence-based. In order to encourage professionals, it would be ideal if most instruments were inexpensive, brief as well as

simple to administer, score and interpret. In addition, any assessment guidelines would need to be concise, straightforward, easily accessible and be regularly updated (Mash & Hunsley, 2005).

Future Directions

Psychological testing and assessment has always been a central component of both clinical and educational psychology (Krishnamurthy et al., 2004). In fact, assessment has become such an integral component of clinical psychology, that it has left the field with the assumption that it is well-equipped to conduct assessments and consequently assessment processes are rarely questioned or evaluated (Fernandez-Ballesteros et al. 2001). However, although there is some empirical support for the use of some anxiety measures used to assess older adults, the majority lack sufficient evidence of their psychometric properties when used with older adults (Therrien & Hunsley, 2012a). More precisely, researchers typically do not analyze, interpret or use reliability data with the same rigor that they do other aspects of their research. Accordingly, I recommend, as have others, that researchers need to routinely report internal consistency reliability coefficients for their own set of data. I also encourage researchers to also present confidence intervals around internal consistency reliability coefficients and confidence intervals for each subgroup. In less ideal cases where coefficients from previous studies are used to conclude that the new scores are reliable, researchers need to explicitly mention that is their deliberate intention and also explicitly defend the use of reliability induction by stating how the two samples are comparable.

For researchers interested in advancing the field of EBA, the situation is definitely one in which the glass can be seen as half full or as half empty. On the positive side, some important initial steps have been taken to implement specific criteria for EBIs and to address the range of issues that need be addressed to further develop the field of EBA. Some useful guidelines and lists of instruments are starting to be available for clinical psychologists to apply in their

assessment activities. Although less is known about the domain of assessment in older adults, the results of the present dissertation can serve as a starting point of which anxiety measure can be considered as evidence-based when used with an older population. This would help in developing and validating evidence-based assessment instruments and protocols against which other anxiety measures could be compared. That being said, much more needs to be accomplished to develop EBA as, at this point, we only have limited consensus on how EBA should be operationalized. Future research efforts need to continue providing evidence of the basic psychometric properties (e.g., reliability and validity) obtained by various measures while increasing the role of incremental validity and clinical utility in the assessment procedures. As assessment is often viewed as a professional skill and clinical activity in its own right, there is a need to re-emphasize the importance of using scientific evidence when selecting assessment instruments and highlight that the interaction between assessment and intervention is at the heart of providing evidence-based psychological services.

References

Achenbach, T.M. (2001). What are norms and why do we need valid ones? *Clinical Psychology: Science and Practice,* 8, 446-450.

Alwahhabi, F. (2003). Anxiety symptoms and generalized anxiety disorder in the elderly: A review. *Harvard Review of Psychiatry, 11*, 180-193.

American Association of Geriatric Psychiatry (2008). Geriatrics and mental health-the facts. Retrieved from http://www.aagponline.org/prof/facts_mh.asp.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological testing*. Washington, DC: American Psychiatric Publishing.

American Psychological Association. (2002). Ethical Principles of Psychologists and Code of Conduct. *American Psychologist, 57*, 1060-1073.

American Psychological Association. (1985). *Standards for educational and psychological Testing.* Washington, DC: American Psychological Association.

American Psychological Association. Division 12. (2004). Guidelines for Psychological Practice in Older Adults. *American Psychologist, 59*, 236-260.

American Psychological Association Presidential Task Force on Evidence Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.

Antony, M. (2002). Enhancing current treatments for anxiety disorders. *Clinical Psychology: Science and Practice, 9*,91-94.

Antony, M., Orsillo, S., & Roemer, L. (2001). *Practionner's guide to empirically based measures of anxiety.* Dordrecht, Netherlands: Kluwer Academic Publishers.

Ayers, C. R., Sorrell, J. T., Thorp, S. R. & Wetherell, J.L. (2007). Evidence-based psychological treatments for late-life anxiety, *Psychology and Aging, 22*, 8-17.

Barlow, D. H. (2002). *Anxiety and its disorders: The nature and treatment of anxiety and panic* (2nd ed.).New York: Guilford Press.

Barnett, D. W., Lentz, F. E. & Macmann, G. M. (2000). Child assessment: Psychometric qualities of professional practice. In E.S Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment* (2nd ed., pp. 355-386). New York: Guilford Press.

Beck, J. G., & Averill, P. M. (2004). Older adults. In R. G. Heimberg, C. L. Turk., & D. S. Mennin (Eds). *Generalized anxiety disorder: Advances in research and practice* (pp.409-433). New York: Guilford Press.

Beck, A. T., Emery, G., & Greenberrg, R. L. (1985). *Anxiety disorders and phobias: A cognitive perspective*. New York: Basic Books.

Beekman, A. T., Bremmer, M. A., Deeg, D. J., van Balkom, A. J., Smit, J. H, de Beurs, E., et al. (1998). Anxiety disorders in later life: A report from the longitudinal aging study Amsterdam. *International Journal of Geriatric Psychiatry, 13*, 717-726.

Bickman, L., Nurcombe, B., Townsend, C., Belle, M., Schut, J., & Karver, M. (1999). *Consumer measurement systems in child and adolescent mental health.* Canberra, ACT: Department of Health and Family Services.

Borkovec, T. D., Robinson, E., Pruzinsky, T., & DePree, J.A. (1983). Preliminary exploration of worry: Some characteristics and processes. *Behaviour Research and Therapy, 21*, 9-16.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1072.

Brown, T. A., & Barlow, D. H. (2002). Classification of anxiety and mood disorders. In D. H. Barlow (Ed.).*Anxiety and its disorders: The nature and treatment of anxiety and panic.*(2nd ed., pp.292-327).New York : Guilford Press.

Bryant, C., Jackson, H., & Ames, D. (2009). Depression and anxiety in medically unwell older adults: Prevalence and short-term course. *International Psychogeriatrics, 21*, 754-763.

Byrne, G. J., & Pachana, N. A. (2011). Development and validation of a short form of the Geriatric Anxiety Inventory-The GAI. *International Psychogeriatrics, 23*, 125-131.

Canadian Study of Health and Aging Working Group. (1994). Canadian Study of Health and Aging: study methods and prevalence of dementia. *Canadian Medical Association Journal, 150*, 899-913

Caruso, J. C. (2000). Reliability generalization on the NEO personality scales. *Educational and Psychological Measurement, 60*, 236-254.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*, 7-18.

Charles, S., Mather, M., & Cartensen, L.L. (2003). Aging and emotional memory: The forgettable nature of negative images for older adults. *Journal of Experimental Psychology, 132*, 310-324.

Clark, D. A., &Beck, A. T. (2010). *Cognitive therapy of anxiety disorders: Science and practice*. New York: Guilord Press.

Clark, L., & Watson, D. (1991). Tripartite model of anxiety and depression psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 3*, 316-336.

Cohen, L. L., La Greca, A. M., Blount, R. L., Kazak, A. E., Holmbeck, G. N., & Lemanek, K. (2008). Introduction to special issue: Evidence-based assessment in pediatric psychology. *Journal of Pediatric Psychology, 33*, 911-915.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test history*. Philadelphia: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J., & Gleser, G.C. (1957). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*,481-489.

Dawson, P., Kline, K., Wiancko, D. C., & Wells, D. (1986). Preventing excess disability in patients with Alzheimer's disease. *Geriatric Nursing, 7*, 298-301.

de Beurs, E., Beekman, A. T., van Balkom, A. J., Deeg, D. J., van Dyck, R., & van Tilburg, W. (1999). Consequences of anxiety in older persons: Its effect on disability, well-being and use of health services. *Psychological Medicine, 29*, 583-593.

Dennis, R. E., Boddington, S. J., & Funnell, N. J. (2007). Self-report measures of anxiety: Are they suitable for older adults? *Aging and Mental Health*, 11, 668-677.

D'Hudson, G., & Saling, L. L. (2010). Worry and rumination in older adults: Differentiating the processes. *Aging and Mental Health, 14*, 524-534.

Diefenbach, G. J., Stanley, M. A., Beck, J. G., Novy, D. M., Averill, P. M., & Swann, A.C. (2001). Examination of the Hamilton scales in assessment of anxious older adults: A replication and extension. *Journal of Psychopathology and Behavioral Assessment, 23*, 117-124.

Dugas,M. J., & Robichaud, M. (2007). *Cognitive-behavioral treatment for generalized anxiety disorder: From science to practice.* New York: Routledge.

Dye, C. (1978). Psychologists' role in the provision of mental health care for the elderly. *Professional Psychology, 9*, 38-49.

Edelstein, B. A., & Segal, D. L. (2011). Assessment of emotional and personality disorders in older adults. In K. Warner Schaie & S. L. Willis (Eds.), *Handbook of the psychology of aging* (7th ed., pp.325-337). San Diego: Academic Press.

Edelstein, B. A., Woodhead, E. L., Segal, D. L., Heisel, M. J., Bower, E. H., Lowery, A. J., & Stoner, S.A. (2008). Older adult psychological assessment: Current instrument status and related considerations. *Clinical Gerontologist, 31(3)*, 1-35.

Elliott, A. N., O'Donohue, W. T. O., & Nickerson, M. A. (1993). The use of sexually anatomically detailed dolls in the assessment of sexual abuse. *Clinical Psychology Review, 13*, 207-221.

Erber, J. T. (2010). *Aging and older adulthood* (2nd edition). Malden, MA: Wiley-Blackwell.

Fernandez-Ballesteros, R., De Bruyn, E. E. J. Godoy, A., Hornke, L. F., TerLaak, J., Vizcarro, C. et al. (2001). Guidelines for the assessment process (GAP): A proposal for discussion. *European Journal of Psychological Assessment, 17*, 187-200.

First, M. B., Pincus, H. A., Levine, J. B., Williams, J. B. W., Ustun, B., & Peele, R. (2004). Clinical utility as a criterion for revising psychiatric diagnoses. *American Journal of Psychiatry, 161*, 946-954.

Fisher, J. E., Goy, E. R., Swingen, D.N., & Szymanski, J. (1994). *Functional characteristics of behavioural disturbances in Alzheimer's disease patients.* Presented at the annual meeting for the Association for Advancement of Behavior Therapy, San Diego, California.

Flint, A. J. (2005). Anxiety and its disorders in late life: Moving the field forward. *The American Journal of Geriatric Psychiatry, 13*, 3-6.

Flint, A. J. (1994). Epidemiology and comorbidity of anxiety disorders in the elderly. *The American Journal of Psychiatry, 151*,640-649.

Forman, S., Fagley, N. S., Steiner, D. D., & Schneider, K. (2009). Teaching evidence-based interventions: Perceptions of influences on use in professional practice in school psychology. *Training and Education in Professional Psychology, 3*, 226-232.

Gatz, M. (1995). *Emerging issues in mental health and aging*. Washington, DC: American Psychological association.

Gatz, M. & Finkel, S. I. (1995). Education and training of mental health service providers. In M. Gatz (Ed.), *Emerging issues in mental health and aging* (pp.282-302). Washington, DC: American Psychological Association.

Gatz, M., & Smyer, M. A. (2001). Mental Health and aging at the outset of the twenty-first century. In J. E. Birren & K. W. Shaie (Eds.), *Handbook of the psychology of aging* (5th edition, pp.523-544). San Diego: Academic Press.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*, 930-944.

Graham, J. M., Diebels, K. J, & Barnow, Z.B. (2011). The reliability of relationship satisfaction: A reliability generalization meta-analysis. *Journal of Family Psychology, 25*, 39-48.

Green, C. E., Chen, C. E., Helms, J. E., & Henze, K. T. (2011). Recent reliability reporting practices in Psychological Assessment: recognizing the people behind the data. *Psychological Assessment, 23*, 656-669.

Gum, A. M., King-Kallimanis, B., & Kohn, R. (2009). Prevalence of mood, anxiety, and substance-abuse disorders for older Americans in the national comorbidity survey replication. *American Journal of Geriatric Psychiatry, 17*, 769-781.

Gurian, B. S., & Miner, J. H. (1991). Clinical presentation of anxiety in the elderly. In C. Salzman, & Leowitz, B. D. (Eds.), *Anxiety in the elderly: Treatment and research* (pp.31-44). New York: Springer Publishing.

Haley, W. E. (1996). The medical context of psychotherapy with the elderly. In S. Zarit & B. Knight (Eds.), *A guide to psychotherapy and aging: Effective clinical interventions in a life-stage context* (pp. 221-239). Washington, DC: American Psychological Association.

Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment, 15*, 456-466.

Helms, J. E., Henze, K. T., Sass, T. L., & Mifsud, V. A. (2006). Treating cronbach's alpha reliability coefficients as data in counselling research. *The Counseling Psychologist, 34*, 630-660.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counselling and Development, 34*, 177-189.

Hinrichsen, G. A., & Dick-Siskin, L. P. (2000). General principles of therapy. In S. K. Whitbourne (Ed.), *Psychopathology in later adulthood* (pp. 323-350). New York: John Wiley.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*, 523-531.

Hunsley, J. & Bailey, J. M. (1999). The clinical utility of the Rorschach: Unfulfilled promises and an uncertain future. *Psychological Assessment, 11*, 266-277.

Hunsley, J., & Lee, C. M. (2010). *Introduction to clinical psychology: An evidence-based approach* (2nd ed.). Toronto: John Wiley & Sons.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3,* 29-51.

Hunsley, J. & Mash E. J. (2008a). *A guide to assessments that work*. New York: Oxford University Press.

Hunsley, J., & Mash, E. J. (2008b). Developing criteria for evidence-based assessment: An introduction to assessments that work. In J. Hunsley & E. J., Mash (Eds.), *A guide to assessments that work* (pp.3-14). New York: Oxford University Press.

Hunsley, J. & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446-455.

Institute of Medicine (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Joint Committee on Testing Practices. (2004). Code of fair testing practices in education. Retrieved from http://www.apa.org/science/FinalCode.pdf.

Kendell, R. E., & Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *American Journal of Psychiatry, 160*, 4-12.

Kessler, K. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*, 593-602.

Kogan, J. N., Edelstein, B. A., & McKee, D. R. (2000). Assessment of anxiety in older adults: Current status. *Journal of Anxiety Disorders, 14*, 109-132.

Knight, B. G. (2004). *Psychotherapy with older adults* (3rd ed.). Thousand Oaks: Ca: Sage.

Knight, B. G., Kaskie, B., Shurgot, G. R. & Dave, J. (2006). Improving Mental Health of Older Adults. In J. E. Birren & K. W. Shaie (Eds.), *Handbook of the Psychology of Aging* (6th ed., pp. 408-419). San Diego: Elsevier Academic Press.

Krishnamurthy, R., Vandecreek, L., Kaslow, N. J., Tazeau, Y. N., Miville, M. L., Kerns, R., et al. (2004). Achieving competency in psychological assessment: Directions for education and training. *Journal of Clinical Psychology, 60*, 725-739.

Kvall, K., Macijauskiene, J., Engedal, K., & Laake, K. (2001). High prevalence of anxiety symptoms in hospitalized geriatric patients. *International Journal of Geriatric Psychiatry, 16*, 690-693.

Lawton, M. P., Kleban, M. H., & Dean, J. (1993). Affect and age: Cross-sectional comparisons of structure and prevalence. *Psychology and Aging, 8*, 165-175.

Lieerman, J. & Stein, M. (2009). Anxiety with comorbid depression : the rule rather than the exception. *Supplement to Psychiatric Times Reporter, Retrieved from www.cmellc.com/images/pdf/cme_supplements/PsychTimes/0904Reporter.pdf.*

Lenze, E. J., Mulsant, B. H., Shear, M.K., Schilberg, H.C., Dew, M. A.Begley, A. E. et al. (2000). Comorbid anxiety disorders in depressed elderly patients. *The American Journal of Psychiatry, 157*, 722-728.

Lenze, E. J., & Wetherell, J. L. (2009). Bringing the bedside to the bench, and then to the community: a prospectus for intervention research in late-life anxiety disorders. *International Journal of Geriatric Psychiatry, 24*, 1-14.

Lobo, A., Launer, L. J., Fratiglioni, L., Andersen, K., Di Carlo, A., Breteler, M. M., et al. (2000). Prevalence of dementia and major subtypes in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. *Neurology, 54, S4-9.*

Lopez-Pina, J., Sanchez-Meca, J., & Rosa-Alcazar, A. (2009). The Hamilton Rating Scale for Depression: A meta-analytic reliability generalization study. *International Journal of Clinical and Health Psychology*, *9*, 143-159.

Lyons, J. S. (2009). Communimetrics: A communication theory of measurement in human service settings. New York: Springer.

McFall, R. M. (2005). Theory and utility-key themes in evidence-based assessment comment on the special section. *Psychological Assessment, 17*, 312-323.

McGrath, R. E. (2001). Toward more clinically relevant assessment research. *Journal of Personality Assessment, 77*, 307-332.

Mash, E. & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Child and Adolescent Psychology, 34*, 362-379.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' response and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Mineka, S., Watson, D., & Clark, L. A. (1998). Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology, 49*, 377-412.

Mischel, W. (1968). *Personality and assessment*. New York: Wiley.

Mitchell, A. J., Rao, S., & Vaze, A. (2010). Do primary care physicians have particular difficulty identifying late-life depression? A meta-analysis stratified by age. *Psychotherapy and Psychosomatics, 79*, 285-294.

Naughton, C., Bennett, K., & Feely, J. (2006). Prevalence of chronic disease in the elderly based on a national pharmacy claims database. *Age and Ageing, 35*, 633-636.

Norcross, J. C., Koocher, G. P., & Garofalo, A. (2006). Discredited psychological treatments and tests: A Delphi poll. *Professional Psychology: Research and Practice, 37*, 515-522.

Norcross, J. C., & Karpiak, C. P. (2012). Clinical Psychologists in the 2010s: 50 years of the APA Division of Clinical Psychology. *Clinical Psychology: Science and Practice, 19*, 1-12.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Onwuegbuzie, A. J., & Daniel, L. G. (2002). A framework for reporting and interpreting internal consistency reliability estimates. *Measurement and Evaluation in counseling and Development, 35*, 89-103.

Pachana, N. A., Byrne, G. J., Siddle, H., Koloski, N., Harley, E., & Arnold, E. (2007). Development and validation of the Geriatric Anxiety Inventory. *International Psychogeriatrics, 19*, 103-114.

Parmalee, P. A. (2007). Depression. In J. E. Birren (Ed.), *Encyclopedia of gerontology: Age, aging and the aged* (2nd ed., pp. 400-409). Boston: Elsevier Academic Press.

Piotrowski, C. (1999). Assessment practices in the era of managed care: Current status and future directions. *Journal of Clinical Psychology, 55*, 787-796.

Qualls, S. H., & Smyer, M. A. (2006). Mental health. In R. Schulz (Ed.), *The encyclopedia of aging* (2nd edition, pp.304-306). New York: Springer.

Riedel- Heller, S. G., Busse, A., & Angermeyer, M. C. (2006). The state of mental health in old-age across the old European Union: A systematic review. *Acta Psychiatrica Scandinavica, 113*, 388-401.

Rousse, S. V. (2007). Using reliability generalization methods to explore measurement error: An illustration using the MMPI-2 PSY-5 scales. *Journal of Personality Assessment*, *88*, 264-275.

Segal, D. L., June, A., Payne, M., Coolidge, F. L., & Yochim, B. (2010). Development and initial validation of a self-report assessment tool for anxiety among older adults: The Geriatric Anxiety Scale. *Journal of Anxiety Disorders, 24,* 709-714.

Seignourel, P. J., Kunik, M. E., Snow, L., Wilson, N., & Stanley, M. (2008). Anxiety in dementia: A critical review. *Clinical Psychology Review, 28*, 1071-1082.

Selignman, L. D., & Ollendick, T. H. (1998). Comorbidity of anxiety and depression in children and adolescents: An integrative review. *Clinical Child and Family Psychology Review, 1*, 125-144.

Shamoian, C. (1991). What is anxiety in older adults. In C. Salzman (Ed.), *Anxiety in the elderly: Treatment and research (*pp. 3-15). New York: Springer.

Sibrava, N. J. & Borkovec, T. D. (2006).The cognitive avoidance theory of worry. In G. C. L. Davey & A. Wells (Eds.). *Worry and its psychological disorders: Theory, assessment and treatment* (pp. 239-256).Chichester, UK : Wiley.

Sinoff, G. & Werner, P. (2003). Anxiety disorder and accompanying subjective memory loss in the elderly as a predictor of future cognitive decline. *International Journal of Geriatric Psychiatry, 18* 951-959.

Stanley, M. A., Beck, J. G., & Zebb, B. J. (1996). Psychometric properties of four anxiety measures in older adults. *Behaviour Research and Therapy, 34,* 827-838.

Statistics Canada. (2005). *What does aging well mean?* Retrieved from http://www.statcan.gc.ca/pub/89-622-x/2006002/4054766-eng.htm.

Streiner, D.L., Cairney, J. & Veldhuizen, S. (2006). The epidemiology of psychological problems of the elderly. *Canadian Journal of Psychiatry, 51*, 185-191.

Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales: A practical guide to their development and use* (3rd edition). New York: Oxford University Press.

Therrien, Z. & Hunsley, J. (2012a). Assessment of anxiety in older adults: A systematic review of commonly used measures. *Aging and Mental Health, 16*, 1-16.

Therrien, Z., & Hunsley, J. (2012b). *Assessment of Anxiety in Older Adults: A Reliability Generalization Meta-Analysis of Commonly Used Measures*. Manuscript submitted for publication.

Thompson, B. & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60(2)*, 174-195.

Turnbull, J.M. (1989). Anxiety and physical illness in the elderly. *Journal of Clinical Psychiatry*, *50*, 40-45.

Vacha-Haase, T. (1998). Reliability generalization exploring variance in measurement error affecting score reliability across studies. *Educational & Psychological Measurement*, *58*, 6-20.

Vacha-Haase, T., Kogan, L. R. & Thompson, B. (2000). Sample composition and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*, 509–552.

Vacha-Haase, T., Ness, C. M., Nilsson, J. & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *Journal of Experimental Education, 67*, 335-341.

Vassar, M., & Bradley, G. (2010). A reliability generalization study of coefficient alpha for the Life Orientation Test. *Journal of Personality Assessment, 9*, 362-370.

Vink, D., Aartsen, M. J., Schoevers, R. A. (2008). Risk factors for anxiety and depression in the elderly: A review. *Journal of Affective Disorders, 106*, 29-44.

Wheeler, D. L., Vassar, M., Worley, J. A., & Barnes, L. L.B. (2011). A meta-analysis of coefficient alpha for the Maslach Burnout Inventory. *Educational and Psychological Measurement, 71*, 231-244.

Wilkinson, L. & American Psychological Association of Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and Explanations. *American Psychologist, 54*, 594-604.

Wisocki, P. A., Handen, B., & Morse, C. K. (1986). The Worry Scale as a measure of anxiety among homebound and community active elderly. *The Behavior Therapist, 9,* 91-95.

Wolitsky-Taylor, K., Castriotta, N., Lenze E. J., Stanley, M. A., & Craske, M. G. (2010). Anxiety disorders in older adults: A comprehensive review. *Depression and Anxiety, 27*, 190-211.

Yates, B. T., & Taub, J. (2003) Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. *Psychological Assessment, 15*, 478-495.

Zarit, S. H., & Zarit, J. M. (2007). *Mental disorders in older adults: Fundamentals of assessment and treatmen*t. New York: The Guilford Press.

Table 1
Summary of Seven Categories of Good Practices for Analyzing, Interpreting, and Using Reliability Data

1. Calculate and report reliability coefficients.
Calculate and report appropriate reliability coefficients for each measure used.
Specify the type of internal consistency coefficients calculated (e.g. Cronbach's alpha, omega, theta, etc.)
Report reliability coefficients and related summary data (e.g. mean scores, standard deviations) for measures as originally designed.
2. Determine whether scale conceptual and structural properties support use of alpha.
Describe theoretical structure of each measure: Is it a multidimensional or multiscale measure? Are responses classically parallel or essentially tau equivalent?
Determine whether measures are indexes or scales. Do not use alpha to describe score reliability of indexes.
3. Engage in researcher scale modifications reluctantly.
When faced with deleting scale items, provide the number of items used in the original and revised measures as well as indices of sample variability and subscale intercorrelations.
4. Describe sample attributes.
Describe sample composition so other researchers can determine whether the reported reliability coefficients are relevant to their samples.
Report sample size to permit inferential statistical analyses.
5. Interpret reliability data.
Conduct reliability generalization analyses to aid in comparing coefficients across studies.
Recognize that obtained alphas are affected by homogeneity of samples.
When reliability coefficients are interpreted, share your analytic process in enough detail for subsequent readers to evaluate your interpretations.
6. Analyze reliability data.
Calculate and report confidence intervals for all reliability coefficients.
If comparing an obtained internal consistency estimate to a referent, use an appropriate statistical coefficient.
7. Use alpha coefficients.
Specify the intended use of the reliability coefficient(s) a priori.
To make inferences about individual's scores, use reliability coefficients to calculate confidence intervals around each person's score(s).
In basic research, correct results for attenuation because of imperfect reliability.

Apendix A

Coding sheet

Coding sheet

Authors :_____
Journal Title :_____
Year :_____
Volume : _____ Page numbers : _____

Does the journal specialize in research on older adults : N
Y

Does the journal specialize in research on the health of older adults? N
Y

Does the journal specialize in research on the mental health of older adults? N
Y

Inclusion criteria :

1. Age:

All participants 65 years and older ? Y N Contact author

All participants 60 years and older ? Y N Contact author

Page number for this information :_____

2. Measures of anxiety :

Type of measure used :

___ Beck Anxiety Inventory
___ Brief Symptom Inventory
___ Clinical Anxiety Scale
___ Depression and Anxiety Stress Scale
___ Geriatric Anxiety Inventory
___ Geriatric Mental State Examination
___ Goldberg Anxiety Inventory
___ Hamilton Anxiety Scale
___ Hospital Anxiety and Depression Scale
___ Mini Mental State Examination
___ Penn State Worry Questionnaire
___ Spielberg State Trait Anxiety Inventory
___ State Trait Anxiety Inventory
___ Structured Clinical Interview (SCID)
___ Other : _____
_____
_____

Score standard deviation: _____ N/A

Page number:_____

First measure used : _____

Does the article indicate that the measure is appropriate for use with older adults? Y N

If yes, does the article indicate it was designed for use with older adults? Y N

If yes, does the article indicate that age-relevant norms are available? Y N

Reliability coefficients reported:

_____

_____

_____

Second measure used (if applicable): _____

Does the article indicate that the measure is appropriate for use with older adults? Y N

If yes, does the article indicate it was designed for use with older adults? Y N

If yes, does the article indicate that age-relevant norms are available? Y N

Reliability coefficients reported:

_____

_____

_____

Third measure used (if applicable): _____

Does the article indicate that the measure is appropriate for use with older adults? Y N

If yes, does the article indicate it was designed for use with older adults? Y N

If yes, does the article indicate that age-relevant norms are available? Y N

Reliability coefficients reported:

_____

_____

Fourth measure used (if applicable): _____

Does the article indicate that the measure is appropriate for use with older adults? Y N

If yes, does the article indicate it was designed for use with older adults? Y N

If yes, does the article indicate that age-relevant norms are available? Y N

Reliability coefficients reported:
_____
_____
_____

Participant variables :

Age range : _____ N/A

Age range for women **:** _____ N/A

Age range for men : _____ N/A

Page number : ____ N/A

Mean age : _____ N/A

Mean age for women : _____ N/A

Mean age for men : _____ N/A

Page number :_____ N/A

Age standard deviation: _____ N/A

Age SD for women: _____ N/A

Age SD for men: _____ N/A

Page number: _____ N/A

Sample:

Where the participants selected because they had:

- a specific medical condition? Y N

Specify:_____
_____

-a specific mental disorder? Y N

Specify:_____
_____

-both? Y N

Specify:_____
_____

From what setting(s) were participants recruited:

- clinical-medical Y N

-clinical-mental health Y N

-community-independent living Y N

-community-residential living Y N

Page number : ___

Number of participants : _____ N/A

Number of women :_____ N/A

Proportion of women: _____ N/A

Number of men : _____ N/A

Proportion of men: _____ N/A

Page number: ___

Appendix B

Coding manual

Coding Manual

DIRECTIVES FOR THE CODING OF ARTICLES

MEASURES OF ANXIETY IN OLDER ADULTS : A META-ANALYTIC RELIABILITY GENERALIZATION STUDY

---

1. Identification material

-*Bibliographical information:* At first, the rater notes the authors, the journal title, the year, the volume and the page numbers of the study in order to have the bibliographical information of every article. This information can normally be found in the first page of the article or in the reference. If ever there is the need to go back to the article, the researchers will be able to do so easily.

-*Journal:* The rater will also need to answer three questions about the source of the article: 1) Does the journal specialize in research on older adult?, 2) Does the journal specialize in research on the health of older adults? and, 3) Does the journal specialize in research on the mental health of older adults? For the first question, the rater circles yes if the journal's main focus is on research in older adults (e.g. American Journal of Geriatric Psychiatry). In the second question, the rater circles yes if the journal does research in the field of health in older adults. The first example (American Journal of Geriatric Psychiatry) would fall in this category. A journal that is specialized in general health (e.g. Journal of Aging and Health) would also fall in this category. Finally, the rater would circle yes for the last question if the journal is specialized in the mental health of older adults. To keep with our first example, The Journal of Geriatric Psychiatry would fall in this category.

2. Inclusion criteria:

Age of participant: The rater needs to answer two questions related to the age of the participant: 1) All participants 65 years and older? and, 2) All participants 60 years and older. Such information is usually found in the résumé of the article or the participant section. The rater circles the appropriate response (yes or no) for each question. If such information cannot be found in the article, the rater needs to circle Contact the author. It is important to note that the rater needs to make sure that *every* participant is above the cutoff age. Having an average age that is above 60 or 65 years is not enough. The rater also needs to note the page in which that information was found

Measure of anxiety: The coding sheet presents a list of the most commonly used measures of anxiety in the general population that was derived from a literature review. The rater has to put a check mark beside *every* measure that was used in the article. If the authors used two different anxiety measures, there should be two checkmarks. If one of the measures used in an article is not present in the list, the rater checkmarks the Other option and writes down the name of the measure.

Reliability coefficients: For every measure that is used in the study, the rater needs to answer three questions: 1) Does this article indicate that the measure is appropriate for use with older adults?, 2) If yes, does the article indicate it was designed for use with older adults? and, 3) If yes, does the article indicate that age-relevant norms are available?. For example, if one article uses the Penn State Worry Questionnaire (PSWQ) and the Stait Trait Anxiety Inventory (STAI), the rater will answer the set of 3 questions for the PSWQ and then for the STAI. The rater circles yes for the first question if the article directly mentions that the anxiety measure is useful to

assess older adults. For the second question, the rater circles yes if the article mentions that the measure was developed specifically for older adults and not for young or middle-age adults. Finally, the rater circles yes for the third question if the article provides statistics on age-relevant norms to prove that the measure is appropriate for older adults. The rater also needs to note, for every reported measure, the alpha coefficients (reliability) *that was reported for the article's data*. Rater will only report coefficient alphas (the most commonly used reliability coefficients).

3. Moderator variables

Age range: The rater notes the age range for the sample as well as the age range for women and men. This information is usually found in the participant section of the article or in a table/figure in the results section. If this information is not reported, the rater circles N/A. The rater also needs to note the page in which that information was found.

Mean age: The rater notes the mean age of the sample as well as the mean age for women and for men. This information is usually found in the participant section of the article or in a table/figure in the results section. If this information is not reported, the rater circles N/A. The rater also needs to note the page in which that information was found.

Age standard deviation: The rater notes the age standard deviation of the sample as well as the age standard deviation for women and for men. This information is usually found in the participant section of the article or in a table/figure in the results section. If this information is not reported, the rater circles N/A. The rater also needs to note the page in which that information was found.

Sample setting: The rater notes if the sample is from a clinical setting, a community setting or a residential setting. The rater chooses clinical setting when the sample consists of older adults with either a general medical problem (e.g. diabetes) or a mental disorder (e.g. generalized anxiety disorder). In this case, the rater also needs to note if the problem is of a physical or mental nature and also specify the disorder. The rater chooses community setting if the sample consists of elders living at home in the community. Finally, the rater chooses residential setting if the sample does not have a specific health problem but lives in a residential home due to old age.

Sample size: The rater notes the sample size as well as the number of men and women participating in the study. This information is usually found in the participant section of the article or in a table/figure in the results section. If this information is not reported, the rater circles N/A. The rater also needs to note the page in which that information was found. It is important to note that the number of participants recruited at the beginning of the study might not be the same as the number of participants who will complete the study. It is therefore important to be careful when noting the sample size.

Percent of males and females: The rater also needs to calculate the proportion of men and women in the study. To do so, the rater needs to divide the number of women by the sample size. The same procedure needs to be used when calculating the proportion of men. This should give a number ranging from 0.0 to 1.0.

Score standard deviation: Finally, the rater notes down the score standard deviation. This information is usually found in the results section. If this information is not reported, the rater circles N/A. The rater also needs to note the page in which that information was found.